SCALABLE SEMI-PARAMETRIC METHODS IN BIOSTATISTICS

by

Detian Deng

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

December, 2017

Abstract

Individualized health, or precision medicine, is an emerging approach for disease prevention and treatment guided by the individual characteristics of the genome, medical imaging, family history, environment and lifestyle of each person. To achieve this goal, it requires efficient and scalable statistical technologies to decipher the connection between this information and the health outcomes. In this thesis, we present statistical methods in support of the goal of individualized health.

In Part I, the primary goal is to provide flexible and efficient estimation to the latent etiology distribution given imperfect measurements. We parameterize the latent etiologic state as a multivariate binary variable, where each binary node represents the presence/absence of an etiologic agent. The multivariate binary measurements are assumed to be conditionally independent given the latent state. Their relation is parameterized by the true positive rates and false positive rates of the measurements. External information extracted from previous literature on the true positive rates are summarized by Beta prior distributions and used to improve the model identifiability. Experts' knowledge on the competition mechanism among

etiologic agents is translated into a sparse correlation structure of the latent state. A scalable Markov Chain Monte Carlo algorithm is proposed for approximating the exact posterior distribution. Also, a variational Bayesian algorithm is developed for fast and even more scalable estimation in case of large-scale problems. We demonstrate the model using the data from the motivating Pneumonia Etiology Research for Child Health (PERCH) study, which aims to provide a comprehensive estimation of the etiology distribution of childhood pneumonia in developing countries.

In Part II, the key objective is to improve the efficiency of survival regression estimators by incorporating external information on the population level survival rates. The accelerated failure time (AFT) model and the Cox proportional hazards model are considered. For each model, the first estimating equation is created based on the benchmark semi-parametric estimator (partial-likelihood estimator for Cox and log-rank estimator for AFT), then additional estimating equations are formed based on the auxiliary survival information. The estimating equations are transformed by applying functional delta method to a set of over-identifying moment conditions. Finally, the parameter estimation and model diagnostics are carried out following the standard generalized method of moments (GMM) framework. We show that the new GMM-based estimators are asymptotically and empirically more efficient than the benchmark estimators. These new estimators are applied to a recent retrospective study on the prognosis of pancreatic cancer.

ABSTRACT

Advisor:

Scott Zeger, PhD

Committee:

Justin Lessler, PhD (chair); Vadim Zipunnikov, PhD; Maria Deloria-Knoll, PhD

Alternates:

Mei-Cheng Wang, PhD; Ravi Varadhan, PhD

Acknowledgments

I am genuinely thankful for the guidance of my advisor, Dr. Scott Zeger, over these years. He has always been patient, supportive and inspiring. He sets a perfect example of a mentor, a statistician, a researcher for me to look up to. The breadth and depth of his knowledge are beyond my imagination. Being an apprentice of his is a precious experience.

I would also like to thank Dr. Chiung-Yu Huang, who advises me on the second part of my thesis. We only worked together for about a year, but our communication has been so effective and smooth. Chiung-Yu is an excellent advisor. Im lucky to have worked with her.

Thanks a lot to Dr. Zhenke Wu, who is a co-advisor to me as well, as we both worked on the PERCH project and he spent a lot of time discussing the problem with me and helped me a lot.

It's my fortune to have Dr. Maria Knoll as my collaborator on the PERCH project. Her scientific insight is exceptional, which makes me think deeply about how I can make a statistical model truly helpful for solving real-world problems.

ACKNOWLEDGMENTS

I would like to thank other members of my defense committee, Justin, Vadim, Mei-Cheng and Ravi for reading my dissertation and providing valuable advice. Its been an enjoyable experience to work with them. Also, thanks a lot to Karen, Brian, Hongkai, and other faculty members for making this department such a wonderful place to be. I feel truly grateful for the past six years I spent here.

Finally, I would like to thank my wife Guan Wang, my parents and friends for their love, encouragement, and support in these years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Statistical challenges in individualized health

For a long time, due to the lack of measurement of key biomarkers and risk factors or the lack of knowledge on how these factors relate to the disease progression, the treatment, and prevention strategies are determined based on the average person with little consideration for the variability across different individuals. However, each person is unique. Although such strategies lead to an overall benefit for the population, an individual's response to the same therapy could vary significantly. In recent years, thanks to the advancement in biomedical and engineering technologies, we can collect and analyze highly complex data, such as DNA sequences, MRI images, and accelerometer signals. It makes individualized health, more recently called "precision medicine", a promising idea for the future of medicine. According to the Personalized Medicine Initiative (Ashley, 2015), it aims to provide disease treatment and prevention based on individual characteristics of the genome, medical imaging, family history, environment, and lifestyle. To help advance this goal from the perspective of statistical learning, we have to overcome challenges in several aspects, which include population etiology estimation, treatment evaluation, and disease prognosis. We will briefly discuss each element in the following paragraphs.

Population disease etiology refers to the distribution of health states in the pop-

ulation. For example, among pneumonia cases, the health state can be defined as the combination of etiologic pathogens that are infecting their lungs. For certain types of cancer, the state may refer to the varied subsets of patients whose cancer was originally caused by different carcinogen exposures. Having a good understanding of the population etiology and its relation to individual covariates is crucial to providing personalized disease prevention plan, such as targeted vaccine access and protective health policy for the high-risk population.

The estimation would be straightforward if we were to observe the health states directly. However, the measurements of these health states are, in many cases, indirect and of varying quality. For example, the Pneumonia Etiology Research for Childhood Pneumonia (PERCH) project (Levine et al., 2012) is a multi-country case-control study to estimate the frequency with which each pathogen causes pneumonia for kids under 5-year old. But sampling directly from a childs lung is not typically feasible given the invasiveness of the procedure. The actual pathogen(s) that infect the lung, therefore, can only be inferred from multiple peripheral measurements, such as nasal swab PCR and blood bacteria culture. The measurements have different sensitivies and specificities. To accurately recover the underlying distribution of the health states in the target population, the statistical estimating procedure has to account for the measurement errors produced in the data collection process.

The primary task in the process of customizing treatment for each patient is to

evaluate the treatment effects in different sub-populations, i.e., sub-group analysis. With the knowledge of the expected treatment effect in each sub-population, clinicians can simply match the individual patient to a sub-group by its profile and adopt the best intervention. A significant challenge in sub-group analysis is the lack of statistical power. For example, there are two types of androgen deprivation therapy (ADT) for men with advanced prostate cancer. Continuous ADT (CADT) is the conventional treatment in the US, and intermittent ADT (IADT) is proposed as an alternative treatment with potential benefits regarding the quality of life, financial cost, and side effects. Moreover, as age and prostate-specific antigen (PSA) level are reportedly the key prognostic factors for advanced prostate cancer, it is of great interest to evaluate the potential advantage of IADT over CADT, and especially to examine whether such potential effect differs by PSA level or age at the time of diagnosis. However, the recent clinical study (Hussain et al., 2013) was not able to prove/disprove the comparative effectiveness of IADT versus CADT in each group of PSA level and age. Therefore, techniques for improving the statistical efficiency are of critical value for sub-group analysis and individualized treatment.

Besides improving the efficiency in sub-group analysis, another task that is crucial to providing individualized treatment is to predict the progress of the disease based on individual biomarkers. Such prediction is especially important for diseases with complex progressions. In those cases, treatment plans must tailor to individual's disease course, and an accurate prognosis will lead to better

treatment response with less harmful side effects. For example, scleroderma is a systemic autoimmune disorder. The pattern of the disease trajectory varies significantly among individuals. Disease complication may take place in various organs at different progression rates to different severity (Varga et al., 2016). Therefore, accurate prediction of the disease trajectory will allow physicians to make individualized treatment plans on which organs should be targeted with aggressive therapy. Besides, personalized cancer therapy (Meric-Bernstam and Mills, 2012) is essentially established on the predictive power of tumor biomarkers on therapy response and disease prognosis. Possible biomarkers include patient genetic factors, tumor molecular characteristics, tumor site as well as demographics. Therefore, predictive modeling with biomarker selection is the critical step that enables the downstream individualized therapies.

1.2 Organizational overview

In this thesis, we present statistical methods in support of the goal of individualized health in two parts. The first part focuses on etiology estimation given imprecise measurements, and the second part focuses on the efficient evaluation of treatment with survival outcomes. In Chapter 2, we first review the relevant statistical models for latent health state estimation, including the latent class model (LCM), grade-of-membership model (GoM), and models for multivariate bi-

nary data. Then we provide an overview of the Bayesian estimating procedures for latent class models. We focus on Markov Chain Monte Carlo (MCMC) techniques and variational Bayesian inference (VBI).

In Chapter 3, we propose a Bayesian latent sparse correlation model for etiology estimation given imperfect measurements. First, we present a brief introduction to the Pneumonia Etiology Research for Child Health (PERCH) study, which aims to provide a comprehensive estimation of the etiology distribution of childhood pneumonia in developing countries. This project is the motivating application of our proposed approach. Furthermore, we introduce a modified quadratic exponential model of the multivariate binary variable with a sparse correlation structure. Then we use it to construct a Bayesian hierarchical model for etiology estimation. A scalable Markov Chain Monte Carlo algorithm with pseudo-likelihood is proposed for posterior approximation. By simulation studies, we show the advantage of our approaches regarding mean estimation error. We also demonstrate the model using the data from PERCH study. In Chapter 4, we provide a fast variational Bayesian algorithm to approximate the posterior of the Bayesian latent sparse correlation model. We show that this approach significantly reduces the estimation time without notable sacrifice in estimation accuracy.

In Chapter 5, we begin by reviewing the statistical methods for improving efficiency by combining auxiliary information. Then we propose a strategy for improving the efficiency of survival regression estimators by incorporating external

information on the population level survival rates. The accelerated failure time (AFT) model and the Cox proportional hazards model are considered. For each model, we describe how we derive the set of over-identifying moment conditions from the benchmark estimators and auxiliary information. Then, the parameter estimation and model diagnostics are carried out following the standard generalized method of moments (GMM) framework. We show that our GMM-based estimators are asymptotically and empirically more efficient than the benchmark estimators. These new estimators are applied to a recent retrospective study on the prognosis factors of pancreatic cancer.

The last chapter concludes this dissertation with its contributions and possible directions for future research.

Chapter 2

Literature Review

2.1 Latent Class Models

When the individual characteristics of interest cannot be directly observed, they are typically represented as latent variables in the statistical models. These models that connect the latent variables to the observables are generally termed as "latent structure models". For example, economists are often interested in studying the quality of life, which cannot be measured directly but can be inferred as a latent variable from the observed attributes such as income, physical and mental health, education, and recreational activities. McCutcheon (1987) categorized the latent structure models into four types, including factor analysis (continuous measurement and continuous latent variables), latent trait model (discrete measurements and continuous latent variables), latent profile model (continuous measurements and discrete latent variables), and latent class model (discrete measurements and discrete latent variables). In this section, we focus on discussing the latent class model (LCM) because in our motivating application, childhood pneumonia etiology estimation, both the measurements and the latent health state are multivariate binary.

In the standard latent class model (Goodman, 1974), the latent variable L is a discrete variable with J mutually exclusive and collectively exhaustive classes. For each individual, we observe K different discrete measurements. With $k = 1, \ldots, K$, the kth measurement variable has D_k categories. Given the latent class

membership, the observed data are assumed to be conditionally independent. Let $\pi_j = P(L = j)$, for $j = 1, \ldots, J$ be the membership probability of each latent class, and define $p_{jkm} = P(M_k = m|L = j)$ as the conditional probability given the jth latent class of observing category m in the kth measurement. The observed data distribution is a finite mixture distribution:

$$P(M_1 = m_1, M_2 = m_2, \ldots, M_K = m_K) = \sum_{j=1}^{J} \pi_j \prod_{k=1}^{K} p_{jkm_k}. \tag{2.1}$$

Applications of LCM can be found in many areas, such as psychiatry (Young, 1983; Sullivan et al., 1998), education (Aitkin et al., 1981) and evaluation of diagnostic tests (Albert et al., 2001). For example, when there is no gold-standard to determine the actual disease status, multiple diagnostic tests are applied to each individual resulting in a set of multivariate binary measurements of the latent disease status. By relying upon the conditional independence assumption, the latent class model is able to estimate 1) the prevalence of each disease state, and 2) the sensitivity and specificity of each diagnostic test. Moreover, by Bayes theorem and results in 1) and 2), we can predict the disease status probabilities for each individual given his/her test results.

A critical issue for the estimation of LCM is the potential non-identifiability. For example, according to Goodman (1974), a LCM with four binary observed variables and three latent classes is not identifiable although there are 15 degrees-of-

freedom with 14 parameters to be estimated. In latent structure models, identifiability is typically discussed in a local sense. Specifically, McHugh (1956) defined that a distribution F is locally identifiable if for the parameter θ_0, there exists a neighborhood $\mathcal{N}(\theta_0)$ such that the following two statements are equivalent: (1) $\forall x$ in the support of F, $F_X(x;\theta_0) = F_X(x;\theta)$ and (2) $\forall\theta \in \mathcal{N}(\theta_0) \cap \Theta$, $\theta = \theta_0$, where X denotes the random vector of observable data and Θ is the parameter space. Given a finite sample size, Berzofsky and Biemer (2012) argued that there are four levels of identifiability rather than a simple dichotomy of identifiable vs. non-identifiable. The first level is identifiable model, where there is a unique globally optimal solution as the information matrix is positive definite. The second is local maxima model, where there are many locally optimal solutions, but it is difficult to identify the global maximum from them. Then the third level is weakly identifiable model, whose likelihood is fairly flat in the neighborhood of a solution (Knott and Bartholomew, 1999). Within this neighborhood, solutions are equally supported by the likelihood. The last one is non-identifiable model, where two or more solutions attain the global maximum, but they are not within the same local region of the parameter space. When the model is not locally identifiable absent additional information, a Bayesian estimator may be identifiable if the prior information favors a single set of parameter variables over the others with equivlent maximized likelihoods. (Lindley, 1972). In practice, when previous knowledge is directly related to the parameters, it is also convenient to adopt the Bayesian framework to overcome

the identifiability problem by directly incorporating previous knowledge into priors, and the resulting posterior distribution naturally combines the prior and likelihood information together (Gustafson, 2009).

When the membership probabilities are considered to be dependent on the covariates, extensions to the standard LCM (Dayton and Macready, 1988), (Bandeen-Roche et al., 1997), and (Huang and Bandeen-Roche, 2004) have been developed to accommodate the regression functionality. These extensions are usually referred to as the latent class regression models (LCRM). The standard LCRM has two key assumptions. The first is the same as for the LCM, that is, given the latent class, measurements are conditionally independent. The other key assumption is that given the latent membership, the covariates are not associated with the measurements. To fit the membership probabilities on the covariate vector X, for $j = 1, \ldots, J$, π_j is re-parameterized as

$$\pi_j = \frac{\exp(\boldsymbol{\beta}_j X)}{\sum_{j'=1}^{J} \exp(\boldsymbol{\beta}_{j'} X)}, \tag{2.2}$$

where $\boldsymbol{\beta}_j$ is the regression coefficient vector for the jth latent class, and $\boldsymbol{\beta}_1$ is typically set to zero since the first class is used as the reference class. Huang and Bandeen-Roche (2004) established the identifiability conditions for the LCRM. In general, the LCRM is identifiable if the standard LCM for the same problem is identifiable and the polytomous logistic regression would be identifiable had the latent

classes been observed. Specifically, when there are K binary measurements and J latent classes, the identifiability requires $2^K - 1 \geq K(J-1)$ and full rank covariate matrix.

2.2 Grade-of-Membership Model

The Grade-of-Membership (GoM) model (Woodbury et al., 1978; Clive et al., 1983) is another important approach for characterizing the distribution of multivariate categorical variables using latent structure. The GoM is also a finite mixture model, but the level of mixture is different from what the LCM has. In the LCM, each realized observation only belongs to one of the latent classes, while the GoM allows an individual to have partial memberships to different classes simultaneously. This partial membership structure has been shown to be useful in many fields, such as genetics (Pritchard et al., 2000) and natural language processing (Blei et al., 2003). Specifically, define $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_J)$, where $\eta_j \in [0, 1]$ for $j = 1, 2, \ldots, J$ and $\sum_{j=1}^{J} \eta_j = 1$ as the latent partial membership vector for an individual. For example, $\eta_{j_1} = 0.5, \eta_{j_2} = 0.5$ with $\eta_{j'} = 0$ for all $j' \neq j_1, j' \neq j_2$, means the individual is half from class j_1 and half from classj_2. Suppose there are K different types of discrete measurements, we define $q_{jkm} = P(M_k = m|\eta_j = 1, \eta_{j'} = 0, \forall j' \neq j)$, where $m = 1, \ldots, D_k$ and D_k is the total number of categories in measurement k. This is the transition probability from the extreme case where the subject belongs to class

j entirely to observing that measurement k has value m. Therefore, the general form of the conditional probability of observing $M_k = m$ given the membership vector $\boldsymbol{\eta}$ is

$$P(M_k = m|\boldsymbol{\eta}) = \sum_{j=1}^{J} \eta_j q_{jkm}, \tag{2.3}$$

Similar to the LCM, the standard GoM also assumes that the measurements $M_1, M_2, \ldots, M_K$ are conditionally independent given the latent membership $\boldsymbol{\eta}$. Denote the joint distribution of $\boldsymbol{\eta}$ as $F_{\boldsymbol{\eta}}$, we can integrate out $\boldsymbol{\eta}$ and get the likelihood function of the observed measurements:

$$P(M_1 = m_1, M_2 = m_2, \ldots, M_K = m_K) \;=\; \int \prod_{k=1}^{K} \left(\sum_{j=1}^{J} \eta_j q_{jkm_k} \right) \mathrm{d}F_{\boldsymbol{\eta}}(\boldsymbol{\eta}). \tag{2.4}$$

Comparing the GoM and the LCM, Manton et al. (1994) showed that the LCM is nested in the GoM structure, while Haberman (1995) argued that the GoM model is a special case of the LCM because by putting certain constraints on the LCM, the resulting distribution of the observed variables is identical to what the GoM specifies. In a brief technical report (Erosheva, 2006) explained how the GoM could be considered as a generalization as well as a special case of the LCM at the same time, and Erosheva et al. (2007) proved the Fundamental Representation Theorem that given K observed variables, any individual-level mixture model (e.g. GoM) with J components can be represented as a constrained population-level mixture model (e.g. LCM) with J^K components.

2.3 Models for Multivariate Binary Data

As in our motivating application, childhood pneumonia etiology estimation, both the measurements and the latent health state are multivariate binary data. We will briefly review the main methods for parameterizing the multivariate binary distribution in this section, and for each method the important parsimonious extensions and corresponding regression models will also be discussed.

- **Multinomial distribution**: Consider a multivariate binary vector of length K, denoted by $L = (L_1, \ldots, L_k)$. There are 2^K possible observations for L, termed cells. Let each cell probability be $P(L = l) = p_l$ with $\sum_l p_l = 1$, then L is a multinomial variable with $2^K - 1$ independent parameters. This is the most straightforward and flexible model but has bad scalability since the number of parameters grows exponentially as the dimension grows. Also, it gives little insight into the structure of the data (Cox, 1972), thus it is hard to find a parsimonious extension of it and few regression models were built upon it.

- **Bahadur representation**: First suggested by Bahadur (1961) and later by Cox (1972), this representation models the joint probability of the multivariate binary data as a functions of the marginal probabilities and the second and higher-order correlation. Let $\theta_j = P(L_j = 1)$ and standardize the data as $U_j = (L_j - \theta_j)/\sqrt{\theta_j(1-\theta_j)}$. Define $\rho_{12\ldots k} = \mathbb{E}(U_1 \ldots U_k)$ as the kth order

correlation between $L_1, \ldots, L_k$. Then the joint probability is defined as

$$P(L = l) = \prod_{j=1}^{K} P(L_j = l_j)\Big\{1 + \sum_{i>j} \rho_{ij} u_i u_j + \sum_{i>j>k} \rho_{ijk} u_i u_j u_k + \ldots + \rho_{12\ldots d} u_1 \ldots u_d\Big\}$$

This representation is also a saturated model with $2^K - 1$ independent parameter. To reduce the number of parameters, one can assume parsimonious models for the correlation structure. For example, one could assume an "exchangeable" correlation structure, in which the kth-order correlations are all the same. Then the parameters would only increase linearly with the dimension. In the extreme case where all correlation parameters are set to zero, this representation becomes an independence model.

Estimation methods for regression models using Bahadur representation were discussed by Lipsitz et al. (1995). Since the maximum likelihood (ML) estimation with a Newton-Ralphston algorithm requires very large sample size compared to the dimension to converge to a unique solution, they proposed the "one-step" ML estimator and proved that it is asymptotically equivalent to the fully iterated ML estimator. An alternative moment-based estimation approach (Lipsitz et al., 1995) was also developed as an extension to Liang and Zeger's (1986) generalized estimating equations (GEE) (Liang and Zeger, 1986).

- **Log-linear Models**: The general log-linear model, first described by Cox

(1972) and discussed in depth by Haberman (1973), is the most widely used parameterization for multivariate binary data. This representation models the joint probability in the log scale as a linear function of conditional log odds' and conditional log odds ratios. It is a member of the exponential family, thus many useful properties can be directly obtained. The general form of log-linear model can be written as:

$$P(L = l; \Theta) = \exp\left\{\Theta_1^T l + \Theta_2^T w_2 + \ldots + \Theta_K^T w_K\right\}/A(\Theta)$$

where w_k is a $\binom{K}{k} \times 1$ vector of the k-way cross-products of l, $k = 1, \ldots, K$, and $\Theta = (\Theta_1, \ldots, \Theta_K)$ contains the the canonical parameters, which is a $(2^K - 1) \times 1$ vector. Θ_1 contains the k conditional log odds' and the rest contains the conditional log odds ratios, regarded as the association parameters. Moreover, let $l^* = (l, w_2, \ldots, w_K)^T$, the normalizing term is defined as

$$A(\Theta) = \sum_{l^*: l \in \{0,1\}^K} \exp\{\Theta^T l^*\}$$

Similar to the Bahadur representation, the above model allows for varying degrees of dependence among $\{L_j\}_{j=1}^K$. Independence model is achieved when all of the tow- and higher-way association parameters are set to zero. And the other extreme is to use the full $2^K - 1$ parameters to form a saturated

model.

A variety of parsimonious extensions and re-parameterizations have been developed based on the log linear model. An important special case is the "quadratic expoential family" described by Zhao and Prentice (1990), which fixes the three- and higher-way association parameters at zero. In addition, they made a one-to-one transformation from (Θ_1, Θ_2) to the marginal moment parameters (μ, σ), where μ is the vector marginal mean and σ is the vector of pairwise covariances, and they derived the likelihood equation for estimating the coefficients of the regression models for μ. However, the problem of this method is that the consistency of the regressions parameters requires the correct specification of both the means and pairwise correlations.

As a method to circumvent the drawback of the above model, an important re-parameterization of the general log-linear model, the "mixed parameter" model, is proposed by Fitzmaurice and Laird (1993). Let $\Omega = (\Theta_1, \ldots, \Theta_K)$, the model is parameterized in terms of (μ, Ω), the mixture of marginal mean and conditional log odds ratios, via the one-to-one transformation from (Θ_1, Ω) to (μ, Ω). Although such transformation has no closed form, the problem can be solved using the iterative proportional fitting algorithm (Deming and Stephan, 1940) within each step of the Fisher scoring algorithm. And it is shown that the regression coefficient estimator is consistent if the mean structure is correctly specified even if the correlation structure Ω is not.

- **Dependence Ratio Model**: The dependence ratio model was proposed by Ekholm et al. (1995), which models the association using dependence ratios rather than odds ratios. Let $\eta = (\eta_1, \ldots, \eta_K, \eta_{12}, \ldots, \eta_{1\ldots K}) = \mathbb{E}(l^*)$. The kth-order dependence ratio is defined as the joint success probability of k binary responses divided by the joint success probability assuming independence. For example, the 2nd order dependence ratio between L_1 and L_2 is $\lambda_{12} = \frac{\eta_{12}}{\eta_1 \eta_2}$. Therefore, dependence ratio being one indicates independence. It is shown that the joint probability can be expressed as an affine linear transformation of η and a marginal regression model is built. Furthermore, Ekholm et al. (2000) suggested five types of parsimonious association models by constraining the structure of η based on this representation.

- **Latent Continuous Distribution**: A multivariate binary distribution can be obtained from a multivariate continuous distribution by thresholding each of the variables. For example, consider a multivariate Gaussian random vector $Z = (Z_1, \ldots, Z_K)$, the corresponding multivariate binary distribution can be constructed by letting $L_j = 1$ if and only if, say, $Z_j > 0$ and letting $L_j = 0$ other wise. This model, considered by Cox (1972), as a "historically important way" and a "useful heuristic device" but "seems unnecessary unless the Z's are of intrinsic interest".

- **Lattice Based Model** The lattice based models are extensively studied and

widely used in the field of spatial analysis and statistical mechanics. The early work can date back to the Ising Model (Ising, 1925) and currently there are two dominant approaches for modeling binary data on a lattice: the spatial generalized linear mixed model which models the dependence by latent Gaussian Markov random field over the lattice (Banerjee et al., 2014) and the autologistic model, which models the dependence directly (Besag, 1974) thorough a linear function of the neighboring variable, termed autocovariate. The later approach is of more interest in terms of our likelihood specification, so we will focus on the autologistic model in this section.

Suppose the multivariate binary data $L \in \{0,1\}^K$ are placed on a lattice. The conditional distribution of L_j is given by:

$$P(L_j|L_{-j}) = \text{logit}^{-1}\Big(\beta_j + \sum_{k\neq j} \alpha_{jk} L_k\Big)$$

where β_j is the conditional log odds, $\{\alpha_{jk}\}$ are the dependence parameters, and the sum is called the autocovariate, which determines the dependence between L_j and all the other variables on the lattice L_{-j}. Let δ_{jk} be the indicator of whether L_j and L_k are neighbors, let D be a $K \times K$ adjacency matrix where $[D]_{jk} = \delta_{jk}$, and assume $\alpha_{jk} = \alpha\delta_{jk}$. By Brook's Lemma, the

joint distribution of L is

$$P(L|\beta,\alpha) = \frac{\exp\left(L^T\beta + \frac{\alpha}{2}L^TDL\right)}{\sum_{Y\in\{0,1\}^K}\exp\left(Y^T\beta + \frac{\alpha}{2}Y^TDY\right)}$$

Thus this model can be also viewed as a special case of the log-linear model.

- **Tensor Factorization Models** The tensor factorization models are essentially latent class models, which assume the multivariate binary outcomes are conditionally independent given one or more latent discrete variables. Marginalization of the latent variables induces the dependence among outcomes and yields a tensor decomposition of the joint probability π. Dunson and Xing (2009) developed a nonparametric Bayes approach using Dirichlet process mixtures of product multinomials, which parametrizes the multivariate unordered categorical outcomes as a single latent class model:

$$\pi = \sum_{h=1}^{H} v_h\Psi_h, \quad \text{with} \quad \Psi_h = \psi_h^{(1)}\otimes\psi_h^{(2)}\otimes\cdots\otimes\psi_h^{(K)},$$

where $\boldsymbol{v} = (v_1,\ldots,v_H)^T$ is the mixing probability vector, ψ_h is the conditional probability vector given the latent class. Dunson and Xing (2009) showed that this model is equivalent to a reduced-rank nonnegative PARAFAC decomposition of the probability tensor π. Bhattacharya and Dunson (2012) extended this approach with a simplex factor model, which can be considered as a mul-

tiple latent class model and induces a nonnegative multilinear singular value decomposition (De Lathauwer et al., 2000), or nonnegative Tucker decomposition (Kim and Choi, 2007), of the joint probability tensor π. The tensor factorization models obtain sparsity in terms of the number of latent factors learned. Moreover, covariate information can be included by stacking the outcome and covariates vector together and modeling this joint vector with the same estimating procedure, then inference on the conditional distribution of the outcome given the covariates is implied by the joint model.

2.4 Markov Chain Monte Carlo Algorithms

A Bayesian approach can overcome the potential non-identifiability in LCM estimation by bringing in prior information (Gustafson, 2009). In our motivating application, the estimation of childhood pneumonia etiology, such prior information is available as the experts' knowledge regarding the quality of measurements. For fitting LCM/LCRM, a Bayesian approach treats the latent variables as additional model parameters, and inference is accomplished by obtaining the joint posterior distribution of all the unknowns (latent variables and parameters). In most cases, this posterior is not available in exact forms and requires approximation. The best known and most popular paradigm to approximate the posterior distribution is to draw samples from it using Markov Chain Monte Carlo (MCMC) techniques.

Markov chain Monte Carlo methods, including Gibbs sampling, Metropolis-Hastings (MH) sampling and extensions, are systematically covered in several books, such as Gilks et al. (1996); Robert and Casella (1999), and Brooks et al. (2011). In general, MCMC techniques have a few significant advantages over other estimating procedures. First, a MCMC algorithm, such as the standard random walk MH sampler (Hastings, 1970), is fairly simple to implement for many complex models, whose parameter inference is hard to obtain by other means. Also, the characteristics of the posterior distribution, such as the posterior means and percentiles, can be inspected easily using the simulated chains. In addition, the posterior of any bounded function of the parameters can be obtained directly by plugging in the simulated values. On the other hand, the major disadvantage of many MCMC algorithms is they are computational intensive and typically cost a lot of time to run until enough effective samples are drawn, especially when the dimension of the parameter space (including latent variables) is high. In addition, MCMC algorithms often require monitoring and tuning to have good mixing and fast convergence. Good mixing means that the Markov chain explores the support of the target distribution thoroughly, and only a small number of iterations are needed between two samples in the chain for them to be approximately independent. Poor mixing often leads to high autocorrelation and very slow convergence. In practice, a popular approach is to visualize the sample path and the autocorrelation function to monitor the chain mixing, and to run multiple chains with different initial values to

check if they converge to the same stationary distribution using the Gelman-Rubin statistic (Gelman et al., 2013). When bad mixing is detected, we need to tune the proposal distribution or re-parameterize (Gelman, 2004) the model to improve chain mixing.

With the parameterization determined, tuning the proposal distribution has two parts. First, we need to make a good choice of the proposal distribution. The standard MH sampler has a symmetric proposal distribution, which is relatively inefficient because it is independent from the model and the data. To circumvent this shortcoming, the Metropolis adjusted Langevin algorithm (MALA) (Roberts and Tweedie, 1996; Roberts and Rosenthal, 1998) and its extensions, such as the manifold MALA (Girolami and Calderhead, 2011), are developed and are shown to provide substantial speed-up of the chain convergence in many applications. Motivated by Langevin diffusions and their equilibrium distributions, the proposal distributions in MALA algorithms make use of the gradient of the log posterior. Therefore, the chain is driven towards regions with high posterior density, which helps improve mixing and accelerate convergence. Roberts et al. (2004) showed that MALA is geometrically ergodic as long as the target distribution is sufficiently smooth, and Bou-Rabee and Hairer (2012) quantified the conditions where MALA could fail at exploring the entire posterior range. Another highly efficient proposal distribution can be found in the Hamiltonian Monte Carlo (HMC) algorithm (Neal et al., 2011), which is motivated by the Hamiltonian dynamics in statistical mechan-

ics. Each HMC iteration starts with simulating an auxiliary variable, the momentum v, from a normal distribution which represents the kinetic energy function. Then the proposal of the new parameter value θ is provided by the leapfrog method, a modified Eulers method to approximate the solution of the Hamiltonian equations on (v, θ). By the nature of Hamiltonian dynamics, the acceptance rate of this proposal step is high. This procedure also avoids using random walks, which decreases the autocorrelation of consecutive samples and leads to sufficient exploration of the target distribution. Approximately solving the Hamiltonian equations requires the gradient of the log posterior density function, thus HMC and MALA algorithms are both restricted to sampling continuous variables and cannot be directly applied to LCM. Gelman et al. (2013) suggested that to fit LCM, we could partition the parameter space into discrete and continuous, then alternate HMC/MALA updates on the continuous parameters and Gibbs or slice updates on the discrete ones.

The second part is to tune the scaling parameters of the chosen proposal distribution. As the tuning procedure could get tedious, adaptive MCMC algorithms are developed to search the optimal tuning parameter value on the fly. The pioneering and most popular work in this field is the adaptive Metropolis (AM) (Haario et al., 2001) algorithm, which adaptively changes the covariance matrix of the normal proposal distribution based on past samples using an efficient recursive formula. Many refined version and extensions of AM are proposed. Among them, the important ones include the delayed rejection adaptive Metropolis (Haario et al., 2006)

and the robust adaptive Metropolis (Vihola, 2012). Moreover, analogs of AM are developed for other MCMC algorithms, such as the adaptive MALA (Marshall and Roberts, 2012) and the adaptive Hamiltonian and Riemann manifold Monte Carlo (Wang et al., 2013). Optimal scaling theories (Roberts, 1998; Roberts and Rosenthal, 1998; Roberts et al., 2001; Beskos et al., 2013) are established to justify and guide the use of these adaptive MCMC algorithms. For example, it is shown that the optimal acceptance rate for MALA algorithms is 0.574 (Roberts et al., 2001) and the step size parameter should be proportional to $d^{-1/3}$ where d is the dimension of the parameter space. These results about the optimal step size also inform the scale of efficiency of each algorithm. The MH algorithms need $O(d)$ iterations to sufficiently explore the state space, MALA algorithms need $O(d^{1/3})$, and the HMC algorithms only need $O(d^{1/4})$ steps, while the actual running time of each step depends on the computational cost of evaluating the density function and its gradient.

2.5 Variational Bayesian Inference

MCMC sampling is widely used for Bayesian inference, but there are still cases that MCMC cannot handle very well. For example, when the data set is extremely large, MCMC algorithms might be too slow in practice. Variational Bayesian inference (VBI) (Jordan et al., 1999; Wainwright et al., 2008) is a good alternative to MCMC sampling when we are in need of a much faster posterior approximation.

In VBI, the main idea is to transform the approximate inference problem into an optimization problem. First, we choose a family of probability densities $\mathcal{Q}$ for the variables of interest $\boldsymbol{\theta}$. Then, the goal is to identify the member in this family that minimizes the Kullback-Leibler (KL) divergence to the true posterior distribution $p(\theta|X)$, that is,

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\text{argmin}}\Big\{\mathbb{E}_q\Big[-\log\frac{p(\boldsymbol{\theta}|X)}{q(\boldsymbol{\theta})}\Big]\Big\}. \tag{2.5}$$

As a result, the minimizer $q^*(\boldsymbol{\theta})$ is the best approximation to the posterior within the variational family $\mathcal{Q}$. Since the marginal distribution of data $p(X)$ is a constant with respect to $q(\boldsymbol{\theta})$, the equation 2.5 is equivalent to

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\text{argmax}}\Big\{\mathbb{E}_q\Big[\log\frac{p(\boldsymbol{\theta},X)}{q(\boldsymbol{\theta})}\Big]\Big\}, \tag{2.6}$$

where $\mathbb{E}_q[\log\frac{p(\boldsymbol{\theta},X)}{q(\boldsymbol{\theta})}]$ is known as the evidence lower bound (ELBO). ELBO can be further transformed into a summation of the expected log likelihood with respect to $q(\boldsymbol{\theta})$ and the negative KL divergence between $q(\boldsymbol{\theta})$ and the prior. Therefore, intuitively, the maximizer of ELBO represents a balance between what is supported by the likelihood and what is favored by the prior.

The choice of the variational family $\mathcal{Q}$ is a key component in VBI. In general, $\mathcal{Q}$ should be flexible enough to describe the characteristics of $p(\boldsymbol{\theta}|X)$ and simple enough so that optimizing the ELBO is efficient. The mean-field family, in which

$q(\boldsymbol{\theta})$ can be represented as a product of mutually independent factors and each factor governs an unique element in $\boldsymbol{\theta}$, is the most popular choice. Extensions of the mean-field family are also developed to allow for interactions between factors. Saul and Jordan (1996) proposed the partially factorized structure for $\mathcal{Q}$, and Barber and Wiegerinck (1999) used the decimatable Boltzmann machine as the variational family. Simulation studies showed that these methods could potentially improve the approximation accuracy, but the computational cost for optimizing the ELBO also increased significantly.

Empirically, many studies have shown that the approximation provided by VBI is reasonably accurate comparing to MCMC in terms of the posterior predictive values (Penny et al., 2003; Blei et al., 2006; Braun and McAuliffe, 2010). But unlike MCMC algorithms, which asymptotically draw samples from the exact target distribution, VBI tends to underestimate the posterior variance. Although there is not much theory developed for the general asymptotic behavior of variational inference, theoretical guarantees of variational approximation have been established for a few individual models and variational families.Wang et al. (2006) showed that the variational posterior mean of the mean-field approximation to the Gaussian mixture model with conjugate prior is consistent. You et al. (2014) showed that in the Bayesian linear model with a normal prior on the coefficients and an inverse gamma prior on the error variance, the variational posterior mean of the mean-field approximation is consistent under standard regularity conditions.

Chapter 3

Bayesian Latent Class Model with Sparse Correlation for Etiology Estimation

CHAPTER 3. BAYESIAN LATENT CLASS MODEL WITH SPARSE CORRELATION FOR ETIOLOGY ESTIMATION

Abstract

Population disease etiology refers to the distribution of health states in the population. Among pneumonia cases, the health state can be defined as the combination of etiologic pathogens that are infecting their lungs. Having a good understanding of the etiology distribution and its relation to individual covariates is crucial to providing personalized disease prevention plan. However, the measurements of these health states are usually indirect, and of varying quality. For example, in the Pneumonia Etiology Research for Childhood Pneumonia (PERCH) project (Levine et al., 2012), sampling directly from a childs lung is not typically feasible given the invasiveness of the procedure. The actual pathogen(s) that infect the lung, therefore, can only be inferred from multiple peripheral measurements, such as nasal swab PCR and blood bacteria culture, with imperfect sensitivity and specificity. In order to accurately recover the underlying distribution of the health states, we propose a modified quadratic exponential model of the multivariate binary latent state with a sparse correlation structure. Then we use it to construct a Bayesian hierarchical model for etiology estimation. A Markov Chain Monte Carlo algorithm with pseudo-likelihood is proposed for posterior sampling. Simulation studies show that our approach provides smaller estimation error than current methods. We also demonstrate the model using the data from PERCH study.

3.1 Introduction

3.1.1 Etiology Study for Childhood Pneumonia

Pneumonia is a form of acute respiratory infection of the lungs. The infection can be caused by a variety of pathogens, indcluding bacteria, viruses, mycobacteria and fungi (Hirama et al., 2011). When a child under five gets pneumonia, the typical symptoms include fever, cough, fast or difficult breathing, lower chest wall indrawing where the chest moves in or retracts during inhalation, and wheezing (Singh and Aneja, 2011). Severe cases may be unable to feed or drink and may also experience unconsciousness, hypothermia and convulsions. Although the majority of child pneumonia cases are nonsevere and can be managed in local primary health care facilities (Levine et al., 2012), the severe/very severe cases may result in death, especially in developing countries. In fact, pneumonia is the single largest infectious cause of death of children under 5 years of age (referred to "children" for the rest part of this article) , with an estimate of 0.92 million deaths per year (as of 2015) accounting for 16% of the total 5.9 million childhood deaths worldwide (Liu et al., 2016; Black et al., 2010). Under the pressure of such a severe public health burden, UNICEF and WHO declared pneumonia to be the "forgotten killer of children" in 2006 (UNICEF et al., 2006) and engaged the Global Action Plan for Prevention and Control of Pneumonia (GAPP) (World Health Organization et al., 2009) in 2009.

CHAPTER 3. BAYESIAN LATENT CLASS MODEL WITH SPARSE CORRELATION FOR ETIOLOGY ESTIMATION

Current prevention and treatment strategies for pneumonia were primarily developed based on the results of early pneumonia etiology studies in the 1980s (Shann, 1986; World Health Organization et al., 1990), in which two bacterial pathogens, streptococcus pneumoniae and haemophilus influenzae, were identified as the primary etiologies of pneumonia mortality. It has been 30 years since those studies conducted, and by 2015, three major changes will have taken place (Levine et al., 2012): the wide use of pneumococcoal and haemophilus influenzae-B conjugate vaccines; the wide spread of HIV infection (Calder and Qazi, 2009); the substantial improvements/changes in living conditions, nutrition, and access to health care. These changes will certainly modify the distribution of pathogens, the transmission, and the natural history of infection, which will make the understandings of pneumonia etiology based on the early studies invalid. Hence the effectiveness of the current prevention and treatment could be greatly diminished. As a result, new information of the current etiology of severe/very severe pneumonia for children under 5 is required to ensure its prevention and treatment strategies are appropriate and effective for the epidemiologic setting of the future. In the context of such a strong need, the Pneumonia Etiology Rearch for Child Health (PERCH) project, the largest of its kind in over 30 years, was launched in 2011 and finished data collection recently.

The PERCH project is a case-control study that enrolled around 9500 children from 7 sites across the globe with three primary goals (Levine et al., 2012). (1) Es-

timate the association between severe/very severe pneumonia and infection with confirmed and putative viral, bacterial, mycobacterial, and fungal pathogens. (2) Learn the probability of severe/very severe pneumonia attributable to each of the candidate pathogens. (3) Evaluate potential risk factors for severe/very severe pneumonia due to novel or under-recognized etiologic pathogens. A case-control design was chosen because it is more efficient than cohort studies and probe studies in terms of identifying the etiology among many different, putative etiologic pathogens. The 7 study sites are in Bangladesh, Gambia, Kenya, Mali, South Africa, Thailand and Zambia. These sites were chosen to represent the developing countries with major childhood pneumonia burdens and a range of diverse epidemiologic settings. The study enrolled about 4200 children hospitalized for severe/very severe pneumonia and approximately 5300 controls randomly selected from the corresponding communities. The inclusion-exclusion criterion are discussed in detail by Deloria-Knoll et al. (2012). For each enrolled subject, data on demographics, known and putative risk factors, and pathogen infection were collected. More explanation on the rationale of the study can be found in the review by Adegbola and Levine (2011).

In order to maximize the detection power and accuracy of pathogen infection, the PERCH investors used multiple specimen types (Hammitt et al., 2012) including blood (for cases only), nasopharyngeal(NP) swab (for both cases and controls), and lung aspirates (for only very few cases). These samples were collected and

tested by a variety of conventional and novel detection techniques such as microscopy, culture, serology, antigen testing, and polymerase chain reaction (PCR) (Murdoch et al., 2012), targeting on more than 30 candidate pathogens.

Tests based on lung aspirates samples are considered to provide the direct observation of the lung and are assumed to have perfect sensitivity and specificity, thus they are called Gold Standard (GS) measurements. Among all peripheral measurements, we assume blood samples provide measurements with perfect specificity, but imperfect sensitivity, and NP samples have both imperfect sensitivity and specificity, thus we call measurements from blood samples Silver Standard (SS) measurements, and those from NP samples Bronze Standard (BS) measurements. For each child (patient) i, let Y_i indicate whether this child is a case ($Y_i = 1$) or a control ($Y_i = 0$). Suppose there are K pre-specified pneumonia causing candidates, the list of measurements can be described by three K-dimensional binary vector: M_i^{GS} (if available), M_i^{SS}, and M_i^{BS}, where $M_{ik}^{Src} = 1$ indicates that the kth pathogen is detected using the $Src \in \{GS, SS, BS\}$ measurements in subject i. The data availability and the format of measurement vector are summarized in Figure 3.1.

3.1.2 Challenges in Statistical Estimation

Due to the invasiveness of the lung aspirate procedure, GS measurements were rarely acquired (Levine et al., 2012; Hammitt et al., 2012). The actual pathogen(s)

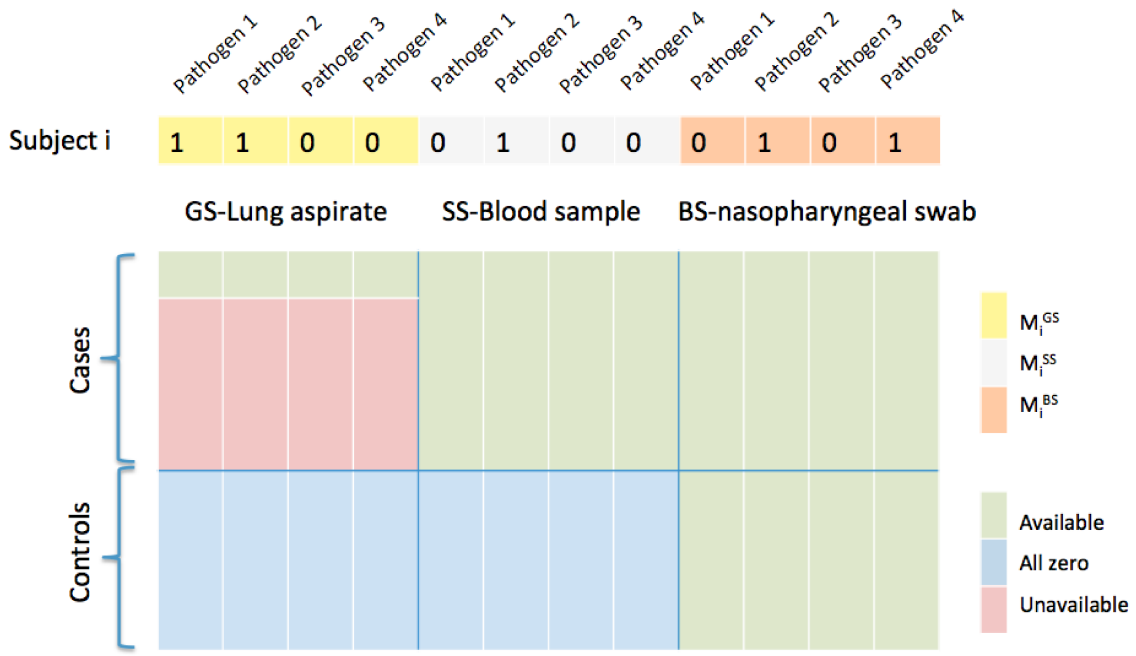

Fig. 3.1: **Data Description**: Suppose we have 4 candidate pathogens in this demonstrative example. In the upper part of this figure, the three 4 dimensional vectors M_i^{GS}, M_i^{SS}, and M_i^{BS} are concatenated together. The GS measurement is available and it tells us pathogen 1 and 2 infect the lung of subject i. Due to the imperfect sensitivity, SS and BS measurements fail to detect pathogen 1. And because of the imperfect specificity, BS measurement detects a false positive for pathogen 4. In the lower part of this figure, the data availability is represented by different colors. As we can see, only a small fraction of cases have GS measurements.

that infect the lung therefore can only be inferred from multiple peripheral measurements with imperfect sensitivity and/or specificity, i.e. the actual lung infection is a latent variable for cases. This fact poses significant statistical challenges for estimating the prevalence of the etiologic pathogens in the population of children, especially in the situation where there are multiple pathogens infecting the lung. Given that most of the measurements are imperfect in terms of sensitivity and specificity, neglecting or inappropriately adjusting (e.g. guessing the wrong value of sensitivity/specificity) for measurement error can produce significantly biased estimates (Gustafson et al., 2002). Therefore, developing a statistical method for estimating μ that appropriately adjusts the measurement errors and incorporates all available sources of evidence is crucial to achieving the goal of PERCH.

Let L_i be a K-dimensional binary vector describing the latent lung infection status for child i, where $L_{ik} = 1$ indicates the child's lung is infected by the kth pathogen. $L_i = (0, \ldots, 0)^T$ means the child has no infection in his/her lung, which is believed to be the true lung status for each control. We also assume, with small probability π_0, a patient identified as case has no infection in his/her lung: $L_i = (0, \ldots, 0)^T$. Using notations defined so far, our interest in this thesis project can be formulated as to estimate the population mean of true lung status of child pneumonia cases given peripheral measurements data, that is, $\mu = \mathbb{E}[L|M^{GS}, M^{SS}, M^{BS}, Y = 1]$. We will call this parameter the etiology fraction in the following discussion.

The latent class model (LCM) (Goodman, 1974) is a statistical model for iden-

tifying unobserved subgroups of the population from multivariate categorical data. The model is parameterized by the prevalence of each latent class and the conditional probabilities for the observed data given each class membership. However, the standard LCM has a few limitations. First, the number of latent classes is determined by comparing the goodness-of-fit of different models; thus the latent classes identified do not always have clear interpretations. Second, even if the set of latent classes were pre-fixed in a meaningful way, the model would suffer from weak model identifiability Goodman (1974) when the number of latent categories is large.

Recently, Wu et al. (2015, 2017) developed the partially-Latent Class Model (pLCM) and the nested-pLCM (npLCM), as extensions to the classic LCM in order to deal with issues mentioned above. In both models, each category of the latent class corresponds to a unique pre-defined state of lung infection. The conditional distributions of the measurements given the latent classes are characterized by sensitivities and specificities. Then the marginal likelihood of the multivariate measurements is modeled as a function of the class prevalence, sensitivities, and specificities. It shows that the full model identifiability can be characterized by the Jacobian matrix of the transformation from model parameters to the distribution of the observables, and in practice, prior scientific information on measurement sensitivities is needed in the absence of GS data. Thus the models are termed partially identifiable (Jones et al., 2010). The limitation of both pLCM and npLCM

is that when K gets large, it is impossible to enumerate every possible infection pattern out of all the 2^K possibilities, while including all 2^K categories will make the computation intractable. To overcome this problem, we propose a novel parameterization of the latent multivariate binary variables and a scalable estimating procedure to handle the high complexity.

3.2 Bayesian Latent Class Model with Sparse Correlation

3.2.1 Parameterization of the Latent Variables

Among various types of parameterization of multivariate binary data we reviewed in section 2.3, the regular quadratic exponential(QE) model described by Zhao and Prentice (1990) is an important special case of the log-linear model, where the three- and higher-way association parameters were fixed at zero. Some general advantages of this model over other parameterization methods include: (1) it is a member of the exponential family; (2) the canonical association parameters are orthogonal to the first order parameters, which is convenient for regression modeling on the conditional mean $\mu = \mathbb{E}(L|M, Z = 1, X)$; (3) the association parameters have intuitive interpretations; (4) the parsimonious assumption is well-accepted and reduces the model complexity to $O(K^2)$. However, under the par-

tially identifiable condition, using regular QE model as the latent parameterization requires informative prior input on all pairwise association parameters. If the prior precision is too low, the parameters will lack identifiability. If the prior precision is made high enough, then when the dimension K gets large, it is practically impossible to find the 'sweet' spot of the prior configuration with $O(K^2)$ hyper-parameters. In this work, we propose a modified quadratic exponential representation for the latent vector L.

Recall that in the pneumonia etiology study example, where each latent binary variable corresponds to a lung infection indicator of a pathogen, different types of pathogens tend to compete with each other during infection (Pericone et al., 2000; Regev-Yochay et al., 2004). Thus these infections are mostly negatively correlated. Motivated by this characteristic, we propose the following Bayesian hierarchical model. Let l be any given realization of the latent vector; let u be the vector of all two-way cross products produced by l; let Θ_1 be the first order canonical parameter vector; let Θ_2 be the second order canonical parameter and $\Theta = (\Theta_1, \Theta_2)$. Then

the probability mass function of the latent vector L is defined as:

$$
\begin{aligned}
P(L = l; \Theta) =& \exp\{\Theta_1^T l + \Theta_2^T u\}/A(\Theta), \qquad (3.1)\\
\text{with } A(\Theta) =& \sum_{l^* \in \{0,1\}^K} \exp\{\Theta^T l^*\}\\
\Theta_2 =& 2\rho \cdot (D_{12}, D_{13}, \ldots, D_{K-1,K})\\
D_{kk'} \sim& \text{Bernolli}(d_{kk'}) \ , \ k, k' \in \{1, 2, \ldots, K\}, k < k'\\
d_{kk'} \sim& \text{Beta}(g_d, h_d),
\end{aligned}
$$

where a single parameter ρ is used to represent the magnitude of the positive/negative association between latent nodes, D_j is the latent indicator of whether the pair of nodes indexed by j are conditionally correlated given all other latent nodes, and D_j has a hierarchical prior distribution with hyper-parameter (g_d, h_d).

This model is motivated by the Bayesian Stochastic Search Variable Selection (George and McCulloch, 1993, 1996; Kuo and Mallick, 1998) method. However, we do not intend to select the association parameter; instead, the mixture posterior distribution of $\rho D_{kk'}$ provides better identifiability in cases of measurement error. Given the resemblance between the proposed model and the classic Bayesian variable selection approach, we name this proposed parameterization as the Sparse Correlation (SC) model, although the association parameters are not necessarily sparse. When covariates are available, stratified estimation for μ can be achieved by parameterizing the first-order canonical parameters for subject i

and latent node k, $\theta_{ik}^{(1)}$, as $X_i^T \beta_k$, where X_i is the vector of covariates. Intuitively, $\theta_k^{(1)}$ stands for the conditional log odds of L_k given all the rest variables, $L_{k'}$ for $k' \neq k$, are zero. Thus, the regression coefficient β_{jk} represents the conditional log odds ratio of L_k associated with one unit increment in covariate $X_{.j}$.

3.2.2 Model Specification

In this section, we derive the full hierarchy of our Latent Sparse Correlation (LSC) Model under three key assumptions: 1) Data are collected using a case-control design, where the latent states of the controls are all zero. 2) GS measurements are not available. 3) SS and BS measurements are conditionally independent given the latent states.

3.2.2.1 The Likelihood Function

Suppose the latent state is of dimension K, let $\gamma \in [0,1]^K$ and $\delta \in [0,1]^K$ represent the True Positive Rate (TPR) and False Positive Rate (FPR) for BS measurements respectively, and let $\eta \in [0,1]^K$ be the TPR for SS measurements, then we have the conditional distribution of the measurements given the latent states.

For the BS measurements:

$$
\begin{aligned}
& P(M_i^{BS}|Y_i, L_i; \gamma, \delta,) \qquad (3.2)\\
&= \prod_{k=1}^{K} P(M_{ik}^{BS}|Y_i, L_i; \gamma, \delta,) \\
&= \exp\Big\{ \sum_{k=1}^{K} \Big[\log(1-\delta_k) + M_{ik}^{BS} \log \frac{\delta_k}{1-\delta_k} + Y_i L_{ik} \log \frac{1-\gamma_k}{1-\delta_k} + Y_i L_{ik} M_{ik}^{BS} \log \frac{\gamma_k(1-\delta_k)}{\delta_k(1-\gamma_k)} \Big] \Big\}.
\end{aligned}
$$

For the SS measurements:

$$
\begin{aligned}
& P(M_i^{SS}|Y_i, L_i; \eta) \qquad (3.3)\\
&= \exp\Big\{ \sum_{k=1}^{K} L_{ik} M_{ik}^{SS} \log \frac{\eta_k}{1-\eta_k} + L_{ik} \log(1-\eta_k) + M_{ik}^{SS} \log L_{ik} \Big\}.
\end{aligned}
$$

Representing the latent distribution function in scalar format, we have

$$
\begin{aligned}
& P(L_i|X_i; \beta, \rho, D) \qquad (3.4)\\
&= \exp\Big\{ \sum_{k=1}^{K} (L_{ik} X_i^T \beta_k + \rho \sum_{k \neq k'} L_{ik} L_{ik'} D_{kk'}) - A(\beta, \rho, D) \Big\},
\end{aligned}
$$

where $A(\beta, \rho, D)$ is the log of the normalizing function. Then the augmented likelihood function is product of term 3.2, 3.3 and 3.4.

3.2.2.2 The Prior Configuration

According to the likelihood function, the parameters of interest are $(\beta, \rho, D, \eta, \gamma, \delta)$. In the joint prior distribution, we assume these parameters are statistically independent. With the control group data available, we can use a Beta$(a'_k = 1, b'_k = 1)$ distribution as the non-informative prior of each δ_k. But for η, γ, informative priors are required. We adopt the same configuration proposed in pLCM Wu et al. (2015), where the joint prior distribution is specified as a product of the individual Beta priors. Let $(\widetilde{a}_k, \widetilde{b}_k)$ be the hyper-parameter that defines the prior of η_k, and let (a^*_k, b^*_k) be the hyper-parameter for γ_k, then the value of $\widetilde{a}_k, \widetilde{b}_k, a^*_k, b^*_k$ are calculated by matching the quantiles of the Beta distributions with the experts' best knowledge on the TPRs and FPRs of each type of measurement. For example, in the Pneumonia etiology research application, the scientists may believe there is 95 $\%$ of chance that the TPR of a SS measurement (e.g. blood culture) is between 0.01 and 0.2, then the value of $(\widetilde{a}_k, \widetilde{b}_k)$ is determined by setting the 2.5th and 97.5th percentile of Beta$(\widetilde{a}_k, \widetilde{b}_k)$ to 0.01 and 0.2 respectively and solving the equation.

For β, the prior is specified as a product of the individual Gaussian distributions, with mean $\mu^*_{\beta_k} = 0$, and variance $\sigma^{*2}_{\beta_k}$. In most cases, we recommend using a relatively large value for the prior variance as we want to stay non-informative for these set of parameters unless strong prior knowledge is present. The prior of ρ is also assumed as a Gaussian distribution, with mean μ^*_ρ, and variance σ^{*2}_ρ, but it requires some extent of prior knowledge to set the values of these two hyper-parameters.

Essentially, the sign of μ_ρ^* defines whether the latent nodes are competing with each other or promoting each other. As defined in formula 3.1, $D_{kk'}$ has a Bernoulli prior with parameter $d_{kk'}$, and each $d_{kk'}$ is a Beta prior that shares the same set of hyper-parameters g_d and h_d.

3.3 Posterior Approximation

Given the model specification in last section, we treat the latent variables as additional model parameters and develop a Markov Chain Monte Carlo (MCMC) algorithm for approximating the posterior distribution of all the unknowns. In general, we partition the parameter space into two parts. For $(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{L}, \boldsymbol{D})$, we apply the Gibbs update based on their full conditionals. For $(\boldsymbol{\beta}, \rho)$ whose exact conditionals are not available, we propose to use the Metropolis adjusted Langevin algorithm (MALA) (Roberts and Rosenthal, 1998; Marshall and Roberts, 2012) updates for fast convergence. Moreover, when sampling $\boldsymbol{D}$ and $(\boldsymbol{\beta}, \rho)$, we avoid calculating the intractable normalizing constant by replacing the true likelihood in equation 4.3 with the following pseudo-likelihood (Besag, 1974).

$$\begin{aligned} &\mathsf{pL}(L_{i1}, L_{i2}, \ldots, L_{iK} | X_i, \boldsymbol{\beta}, \rho, \boldsymbol{D}) \\ &= \exp\Big\{ \sum_{k=1}^{K} \big[L_{ik} X_i^T \beta_k - \log\big(1 + \exp(X_i^T \beta_k + \rho \sum_{k' \neq k} L_{ik'} D_{kk'})\big)\big] \Big\} \end{aligned} \tag{3.5}$$

Next, we present the full conditionals of $(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{L}, \boldsymbol{D})$ for Gibbs updates.

• $\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}$

Based on the conditional probability functions in 4.1 and 4.2, we can show that the full conditionals of $\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}$ are independent Beta distributions. For $k = 1, 2, \ldots, K$, $p(\eta_k|\boldsymbol{M}, \boldsymbol{L}) = \text{Beta}(\widetilde{A}_k, \widetilde{B}_k)$, $p(\gamma_k|\boldsymbol{M}, \boldsymbol{L}) = \text{Beta}(A_k^*, B_k^*)$, and $p(\delta_k|\boldsymbol{M}, \boldsymbol{L}) = \text{Beta}(A_k', B_k')$. We assume that the first n_{case} observations are from the case group, and denote the total sample size as n, then

$$\begin{aligned}
\widetilde{A}_k &= \sum_{i=1}^{n_{case}} L_{ik} M_{ik}^{SS} + \widetilde{a}_k \\
\widetilde{B}_k &= \sum_{i=1}^{n_{case}} L_{ik}(1 - M_{ik}^{SS}) + \widetilde{b}_k \\
A_k^* &= \sum_{i=1}^{n} Y_i L_{ik} M_{ik}^{BS} + a_k^* \\
B_k^* &= \sum_{i=1}^{n} Y_i L_{ik}(1 - M_{ik}^{BS}) + b_k^* \\
A_k' &= \sum_{i=1}^{n} M_{ik}^{BS}(1 - Y_i L_{ik}) + a_k' \\
B_k' &= n + \sum_{i=1}^{n} (Y_i L_{ik} M_{ik}^{BS} - Y_i L_{ik} - M_{ik}^{BS}) + b_k'.
\end{aligned} \tag{3.6}$$

• $\boldsymbol{L}$

In recent approaches (Wu et al., 2015, 2017), the latent multivariate binary vector is updated collectively as one multinomial variable with up to 2^K unique categories. This method is apparently intractable with large K, thus we propose to update individual binary state sequentially. For $i = 1, 2, \ldots, n_{case}$ and $k = 1, 2, \ldots, K$, the

full conditional of L_{ik} is a Bernoulli distribution with probability π_{ik}, such that

$$\pi_{ik} = \frac{\exp(H_{ik} + \rho \sum_{k' \neq k} L_{ik'} D_{kk'})}{1 + \exp(H_{ik} + \rho \sum_{k' \neq k} L_{ik'} D_{kk'})}, \tag{3.7}$$

where $H_{ik} = X_i^T \beta_k + \log \frac{1-\gamma_k}{1-\delta_k} + M_{ik} \log \frac{\gamma_k(1-\delta_k)}{\delta_k(1-\gamma_k)}$.

- $\boldsymbol{D}$

For $k_1, k_2 = 1, 2, \ldots, K$ and $k_1 < k_2$, the full conditional of $D_{k_1 k_2}$ is also a Bernoulli distribution, such that

$$\begin{aligned} p(D_{k_1 k_2}|\cdot) \propto \exp\Big\{ \sum_{i=1}^{n_{case}} \big[& 2\rho L_{k_1} L_{k_2} D_{k_1 k_2} - \\ & \log(1 + e^{X_i^T \beta_{k_1} + \rho \sum_{k \neq k_1, k \neq k_2} L_{ik} D_{kk_1} + \rho L_{ik_2} D_{k_1 k_2}}) - \\ & \log(1 + e^{X_i^T \beta_{k_2} + \rho \sum_{k \neq k_1, k \neq k_2} L_{ik} D_{kk_2} + \rho L_{k_1} D_{k_1 k_2}}) \big] + \\ & D_{k_1 k_2} (\log \frac{d_{k_1 k_2}}{1 - d_{k_1 k_2}}) \Big\} \end{aligned} \tag{3.8}$$

We denote the full conditionals of $(\boldsymbol{\beta}, \rho)$ as $\pi(\boldsymbol{\beta}, \rho)$, then

$$\begin{aligned} \pi(\boldsymbol{\beta}, \rho) \propto \exp\Big\{ & \sum_{i=1}^{n_{case}} \sum_{k=1}^{K} \big[L_{ik} X_i^T \beta_k - \log\big(1 + \exp(X_i^T \beta_k + \rho \sum_{k' \neq k} L_{ik'} D_{kk'})\big) \big] - \\ & \sum_{k=1}^{K} \frac{(\beta_k - \mu_\beta^*)^2}{2\sigma_\beta^{*2}} - \frac{(\rho - \mu_\rho^*)^2}{2\sigma_\rho^{*2}} \Big\} \end{aligned} \tag{3.9}$$

Let $\psi = (\boldsymbol{\beta}, \rho)$ and assume it has dimension d. Define ϵ_t as a random draw from

$N(0, I_{d\times d})$ at the tth iteration. The MALA proposal for the next iteration is

$$\psi^{(t+1)} = \psi^{(t)} + \frac{h}{2}\nabla \log \pi(\psi^{(t)}) + \sqrt{h}\epsilon_t, \tag{3.10}$$

where h is the tuning parameter. The Metropolis-Hastings ratio α is

$$\alpha = \frac{\pi(\psi^{(t+1)})q(\psi^{(t)}, \psi^{(t+1)})}{\pi(\psi^{(t)})q(\psi^{(t+1)}, \psi^{(t)})}, \tag{3.11}$$

where $q(\psi', \psi) \propto \exp\{-\frac{1}{2h}||\psi' - \psi - \frac{h}{2}\nabla \log \pi(\psi)||_2^2\}$. Then with probability $\min(1, \alpha)$, we accept the proposal $\psi^{(t+1)}$, otherwise, keep $\psi^{(t+1)} = \psi^{(t)}$. According to Roberts et al. (2001), we adjust the value of h so that the overall acceptance rate is around 0.57.

3.4 Simulation Study

3.4.1 Design of Study

Three sets of simulation studies are carried out to empirically evaluate the effectiveness of the LSC model under different situations. For all three sets of studies, we assume there are five candidate pathogens ($A, B, \ldots, E$) and two relevant binary covariates (X_1 and X_2). In each study, 200 data sets are simulated independently, and in each simulated data set, there are 500 case subjects and 1000

control subjects.

At the data simulation stage, we first simulate the true lung infection status, then generate the BS and SS measurements. In Study I, multiple pathogens are allowed to infect the lung at the same time, and the measurements are of relatively low quality, that is, lower true positive rates and higher false positive rates. In Study II, infection is assumed to be caused by a single pathogen, and the measurement quality is the same as in Study I. In Study III, the actual lung status is generated in the same way as in Study I, but the measurements have relatively high quality. Details of the study design are described below.

I In the first set of studies, the true lung infection status L of case patients are generated by a Quadratic Exponential Model, where the first order canonical parameters are dependent on both covariates with an interaction effect, and the second order parameters are independent of the covariates. For case subject i and the kth pathogen,

$$\theta_{ik}^{(1)} = \beta_{k0} + \beta_{k1}x_{i1} + \beta_{k2}x_{i2} + \beta_{k3}x_{i1}x_{i2}$$

Also, with $K = 5$, there are 10 second order parameters. We assume that two of them are zero, which represents that two particular pairs of pathogens ($B : C$ and $B : D$) infect lungs independently from each other. The rest eight association parameters share the same negative value: -1.5, which

stands for the pairwise competition among pathogens. Then, the BS and SS measurements for case subjects are simulated based on formula (8) and (9) respectively, and the SS measurements for control subjects are simulated based on formula (10) assuming that there is no infection at all in control patients' lungs, where $\text{TPR}^{(SS)} \approx 0.1, \text{TPR}^{(BS)} \approx 0.7, \text{FPR}^{(BS)} \approx 0.45$. The actual parameter values used in the simulation process are summarized in table 3.1.

Table 3.1: The model parameters used for data simulation in study I

Pathogen	A	B	C	D	E
β_0	0.21	-0.28	-0.84	-0.21	1.07
β_1	-0.1	-0.5	0.5	0.2	0.1
β_2	-0.3	0.2	-0.2	-0.1	0.3
β_3	0.4	0.3	-0.4	0.2	-0.2
$\text{TPR}^{(SS)}$	0.11	0.12	0.08	0.15	0.10
$\text{TPR}^{(BS)}$	0.80	0.60	0.70	0.70	0.65
$\text{FPR}^{(BS)}$	0.50	0.55	0.40	0.35	0.45

II In the second set of studies, the true lung infection status L of case patients are generated by a Multinomial Model, which is equivalent to the above QE model but with all the second order parameters set to negative infinity. The multinomial etiology probabilities are listed in table 3.2. Also, the BS and SS measurements are generated in the same way as they are in Study I with the same TPRs and FPRs.

III In the third set of studies, the true lung infection status of case patients are simulated from the same model as in study I, but the parameters that control

Table 3.2: The etiology probabilities used for data simulation in study II

	Other	A	B	C	D	E
strata 1	0.200	0.241	0.079	0.088	0.125	0.267
strata 2	0.171	0.224	0.113	0.106	0.191	0.196
strata 3	0.232	0.191	0.049	0.115	0.105	0.308
strata 4	0.224	0.176	0.072	0.137	0.162	0.230

the measurement quality are set differently, that is $\text{TPR}^{(SS)} \approx 0.8, \text{TPR}^{(BS)} \approx 0.9, \text{FPR}^{(BS)} \approx 0.05$. Their values are listed in table 3.3.

Table 3.3: The model parameters used for data simulation in study III that are different from study I

Pathogen	A	B	C	D	E
$\text{TPR}^{(SS)}$	0.81	0.82	0.88	0.85	0.80
$\text{TPR}^{(BS)}$	0.98	0.96	0.97	0.97	0.95
$\text{FPR}^{(BS)}$	0.050	0.055	0.040	0.035	0.045

In each of the above situation, 200 independent data sets are generated. The LSC model is applied to each data set (without using GS measurements) with a series of different prior specified by μ_ρ^*, g_d and h_d, which represent the experts' prior knowledge on the magnitude of the competitions between pathogens. Within each study, the values for hyper-parameters $\widetilde{a}_k, \widetilde{b}_k, a_k^*, b_k^*, a_k'$ and b_k', $k = 1, 2, \ldots, 5$, which control the prior input on measurement quality, do not vary. Wu et al. (2015) had discussed the model sensitivity to these hyper-parameters and the partial identifiability issue, which also applies to our method. Thus we do not further study the sensitivity issue on these hyper-parameters. Their values are selected according to experts' knowledge on the quality of BS and SS measurements, and the true

TPR and FPR values are set to be covered by the prior 95% credible interval. For all three studies, same values are used for σ_ρ, and σ_β, which are set large enough to represent non-informativeness. These hyper-parameter values are listed in table 3.4.

Table 3.4: The common hyper-parameters used for model fitting in simulation studies

	$\widetilde{a}_k$	$\widetilde{b}_k$	a_k^*	b_k^*	a'_k	b'_k	σ_β	σ_ρ
Study I	7.6	59	12.7	4.8	1	1	2.2	2.2
Study II	50	10	12	1	1	1	2.2	2.2
Study III	7.6	59	12.7	4.8	1	1	2.2	2.2

For each different prior specification in each study, we have 200 sets of posterior samples produced by the LSC model. Their posterior means are collected to construct an approximate sampling distribution of the estimator. The average of these approximate sampling distributions implies empirically the values to which our parameter estimates converge. Thus the overall accuracy of the LSC model is evaluated based on these sampling distributions means. Note that a five-dimensional multivariate binary distribution can be represented by a multinomial distribution with 32 cells. Let q_j, $j = 1, \ldots, 32$ be the true multinomial cell probabilities, and let $\widehat{q}_j$ be the cell probability estimations based on the sampling distribution means, then the Bhattachayya coefficient (Bhattachayya, 1943), $\sum_{j=1}^{32} \sqrt{q_j \widehat{q}_j} \in [0, 1]$, which measures the similarity between two discrete distributions, is a good metric of the general accuracy of the LSC model.

In Study I and II, the LSC model is compared against the Bayesian partially-

Latent Class Model (pLCM) Wu et al. (2015). The pLCM, which originally only considers single-pathogen infection, now allows multi-pathogen infection in its latest release by making it possible to manually specify candidate pathogen combinations and a Dirichlet prior with equal weights. However, the saturated model, in which all possible combinations are included, is unstable and lack of identifiability with only a few hundred case subjects. Thus, rather than the saturated model, two most commonly used pLCM specifications are applied: 1) the classic pLCM (pLCM-1) where only single pathogen infections are allowed and 2) the new pLCM (pLCM-2) where not only single pathogen but also all pairs of pathogen infections are allowed.

3.4.2 Results

Table 3.5 lists the Bhattachayya coefficients of the LSC model under different prior specifications, and of the two pLCMs in Study I. The same information is visualized in figure 3.2. As we can see, when the true data generating mechanism allows multi-pathogen infection, the classic pLCM (pLCM-1) performs the worst, the pLCM-2 is the second worst, and the LSC model, across all prior specifications, shows a significant amount of improvement over the two pLCM models. The variations caused by different prior specifications is relatively small comparing to the improvement over pLCM models, especially the single-pathogen model.

Model performances in Study II is summarized in table 3.6 and plotted in figure

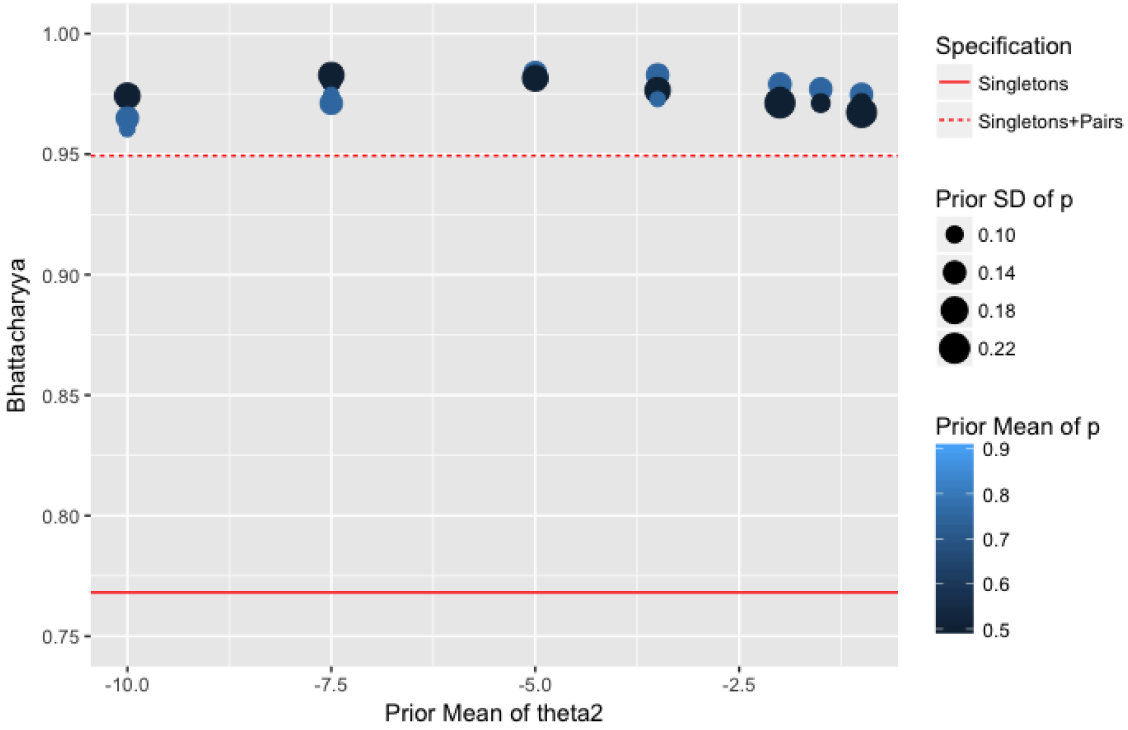

Fig. 3.2: **Overall estimation accuracy of models in study I**: In this study, the data is generated from a quadratic exponential model with low measurement quality. The x-axis stands for μ_ρ^*, the prior mean of ρ, and the y-axis represents the Bhattacharyya coefficient (BC). The round dots are values generated by LSC models with different priors. The color of dots indicates the mean of the prior distribution of p and the size of dots indicate the standard deviation of the prior distribution of p. These two values are calculated based on g_d and h_d. The two horizontal lines are benchmark values generated by pLCM models. The solid line represents the pLCM-1 model which only allows singleton infections, and the dashed line represents the pLCM-2 model which allows singleton and all the pair infections.

Table 3.5: Summary of the overall parameter estimation accuracy of each model fitted in study I

μ_ρ^*	g_d	h_d	Bhattacharyya
-5	6	2	0.9838
-3.5	6	2	0.9828
-7.5	4	4	0.9828
-5	4	4	0.9815
-2	6	2	0.9790
-2	15	5	0.9790
-1.5	6	2	0.9770
-3.5	4	4	0.9765
-1	6	2	0.9748
-7.5	15	5	0.9747
-10	4	4	0.9742
-3.5	15	5	0.9728
-2	2	2	0.9712
-7.5	6	2	0.9711
-1	15	5	0.9677
-1	2	2	0.9673
-10	6	2	0.9649
-10	15	5	0.9604
pLCM-2			0.9494
pLCM-1			0.7681

3.3. The figure shows that when the true data generating mechanism is single pathogen infection only, the pLCM-1 has the best estimation accuracy since that is the true model, and the pLCM-2 performs the worst because it tends to attribute the cause of the disease to two-pathogen infection. The LSC model, on the other hand, shows an increasing trend in estimation accuracy as the prior mean of θ_2 gets more negative because $\theta_2 = -\infty$ makes the model equivalent to the true model. Moreover, in practice, as we can see in figure 3.3, if the true model is not known, the LSC model can still provide an estimate almost as accurate as the

pLCM-1 with a moderately negative (e.g. between -10 and -5) prior mean of θ_2.

Table 3.6: Summary of the overall parameter estimation accuracy of each model fitted in study II

$\mu^*\rho$	g_d	h_d	Bhattacharyya
-16	6	2	0.9925
-10	6	2	0.9828
-16	4	4	0.9761
-7.5	15	5	0.9742
-10	15	5	0.974
-7.5	6	2	0.9725
-10	4	4	0.9717
-16	15	5	0.9635
-7.5	4	4	0.9603
-4	6	2	0.9501
-4	15	5	0.9257
-1	15	5	0.9101
-1	6	2	0.9089
-1	4	4	0.9086
-4	4	4	0.9052
pLCM-1			0.9981
pLCM-2			0.8203

It is very intuitive to think that better measurement quality leads to more accurate etiology estimation. Figure 3.4, which compares the etiology estimations, $\widehat{\mathbb{E}}(L|Y=1,X)$, from study III and two scenarios of study I, confirms this hypothesis. In this figure, the sampling distributions of the etiology estimations for pathogen A, B, and C in the upper panel have larger standard deviations than those in the middle panel. Also, their shapes are more skewed, and the estimation bias tends to be larger. The distributions for pathogen D and E are almost the same in the upper and the middle panel. Comparing the lower panel to the middle panel, we can see that the sampling distributions estimated from high-quality measurements

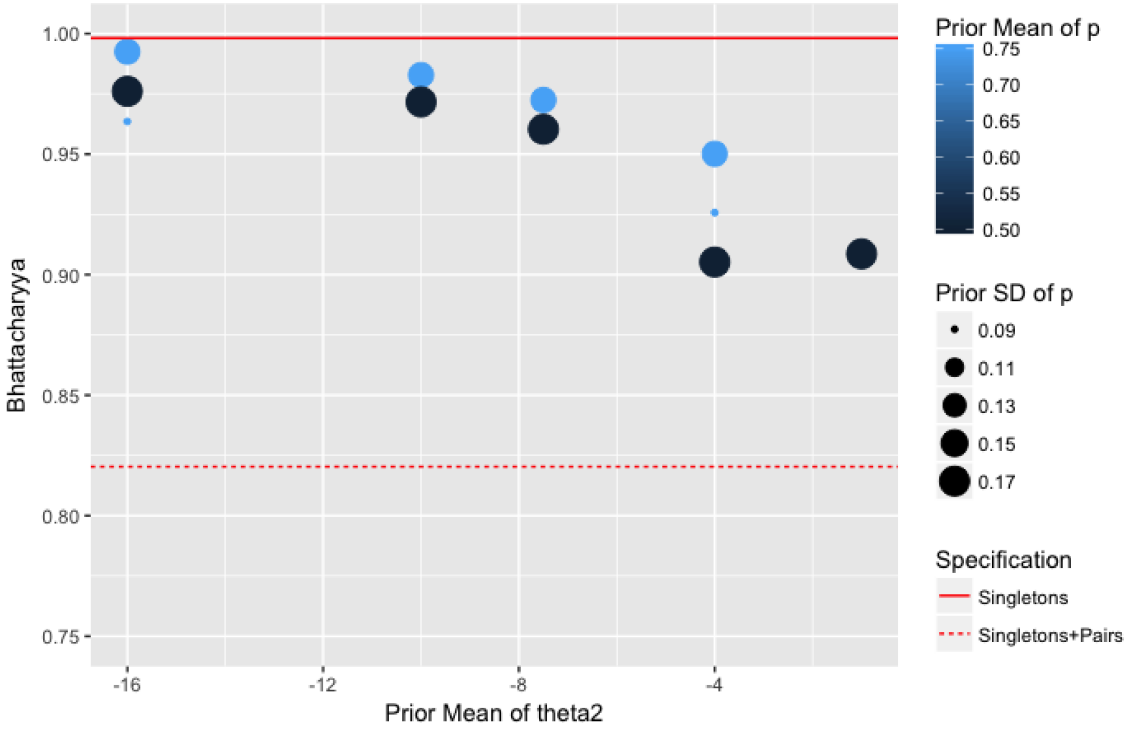

Fig. 3.3: **Overall estimation accuracy of models in study II**: In this study, the data is generated from a singleton-only model with low measurement quality. The x-axis stands for μ_ρ^*, the prior mean of ρ, and the y-axis represents the Bhattacharyya coefficient (BC). The round dots are values generated by LSC models with different priors. The color of dots indicates the mean of the prior distribution of p and the size of dots indicate the standard deviation of the prior distribution of p. These two values are calculated based on g_d and h_d. The two horizontal lines are benchmark values generated by pLCM models. The solid line represents the pLCM-1 model which only allows singleton infections, and the dashed line represents the pLCM-2 model which allows singleton and all the pair infections.

have much smaller standard deviations and nearly no estimation bias.

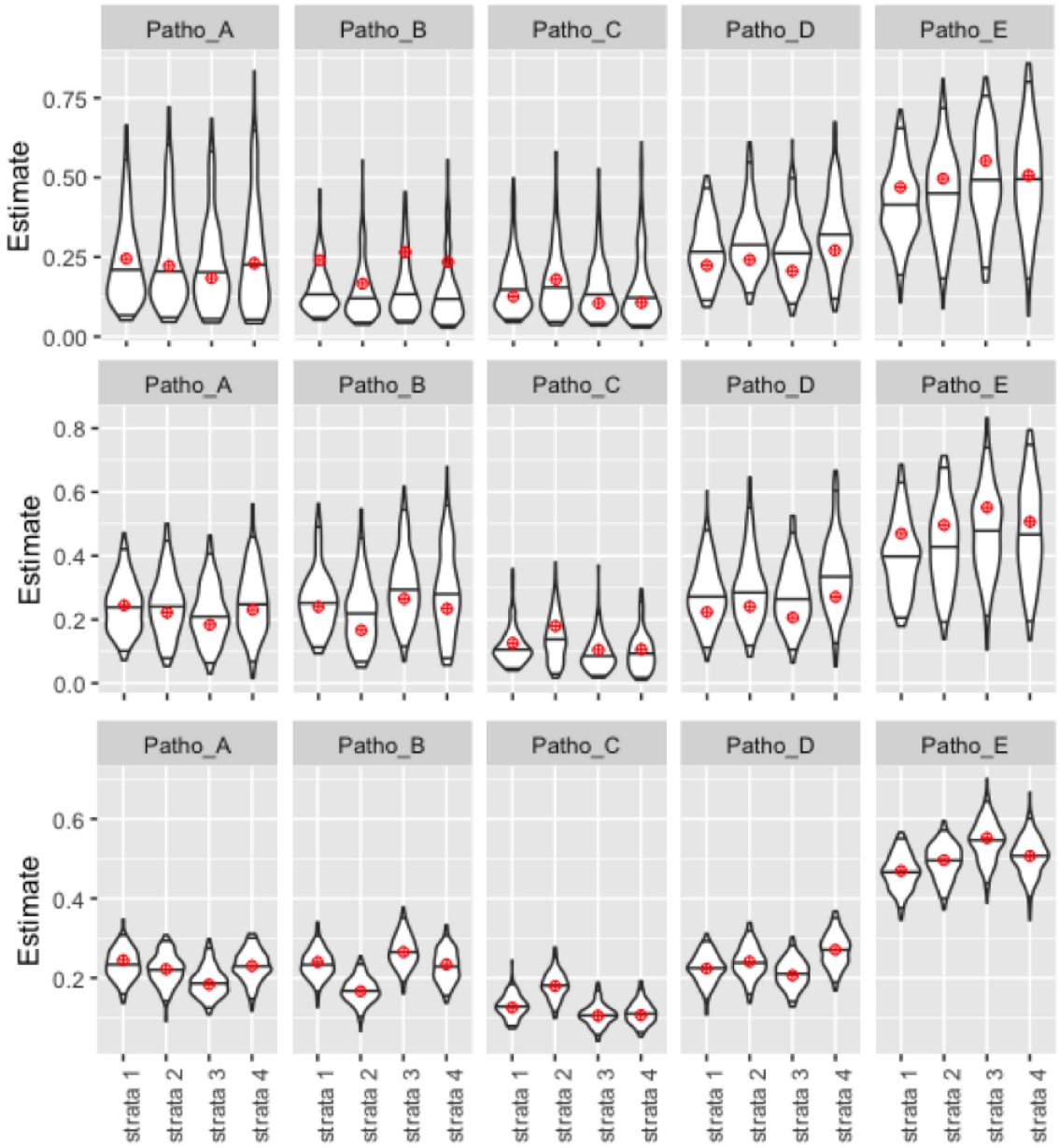

Fig. 3.4: **The impact of data quality on the LSC model**: The results from Study I, where the data is generated from a quadratic exponential model with low measurement quality and with SS measurements available for all pathogens, are plotted in the middle panel. The results from Study I with SS measurements of pathogen A, B, and C removed are plotted in the upper panel. The results from study III, where the data is generated from the same quadratic exponential model as in study I with high measurement quality, are plotted in the lower panel. In each plot, there are five facets, with each one corresponds to a pathogen. In each facet, the x-axis represents four strata that are determined by the two binary covariates, and the y-axis stands for the estimated value of etiology probability, $\widehat{\mathbb{E}}(L|Y=1,X)$. The violin shape indicates the estimated density function of the sampling distributions of $\widehat{\mathbb{E}}(L|Y=1,X)$. The three horizontal lines in each violin shape represent the $2.5th$, $50th$, and $97.5th$ percentiles respectively. The red dots show the true values of the corresponding parameters.

Besides the marginal etiology probability $\mathbb{E}(L|Y=1,X)$, we also compare the estimated probabilities of the most prevalent etiological combinations of pathogens. Figure 3.5 shows the sampling distributions of the singleton and doubleton etiology probabilities estimated in three scenarios. The first thing we can learn from the figure is that in the ideal circumstance with high-quality measurements, the LSC model can provide accurate etiology probability estimates for every single combination of pathogens. While in reality, where measurement quality is relatively low and multiple pathogens do not have SS measurements, the LSC model provides more accurate (less bias and smaller variance) estimate for most etiology combinations than the pLCM-2 model. A general observation is that the pLCM-2 estimates tend to over-estimate the doubleton probabilities. Our explanation is that the multinomial likelihood in the pLCM-2 model does not take the interaction structure into account nor does it provide shrinkage on the probability estimates. Thus it attributes some

of the singleton or tripleton combinations to doubleton combinations.

3.5 Analysis of PERCH Data

The PERCH study enrolled about 4200 children hospitalized for severe/very severe pneumonia and approximately 5300 controls randomly selected from communities across 7 sites around the world. To demonstrate the application of the LSC model for the analysis of PERCH study data, only the Kenya site data, where there is good availability of both BS and SS measurement data, is used so that the site-specific effect is not a concern. We picked the top 5 pathogens reported in Wu et al. (2015) as our candidate pathogens in this analysis. These pathogens are streptococcus pneumoniae (PNEU), haemophilus influenzae (HINF), human metapneumovirus type A or B (HMPV_A_B), rhinovirus (RHINO), and respiratory syncytial virus type A or B (RSV). The BS measurements (nasopharyngeal specimen with PCR detection of pathogens - NPPCR) are available for all 281 cases and 1138 frequency-matched controls on all 5 pathogens. The SS measurements (blood culture results - BCX) are only available for all cases on the two bacteria pathogens: PNEU and HINF.

Prior scientific knowledge Murdoch et al. (2012) suggests that the TPR of BS measurements (NPPCR) is in the range of 50% - 99%, and the TPR of SS measurements (BCX) is in the range of 5% - 20%. Thus we set the hyper-parameters

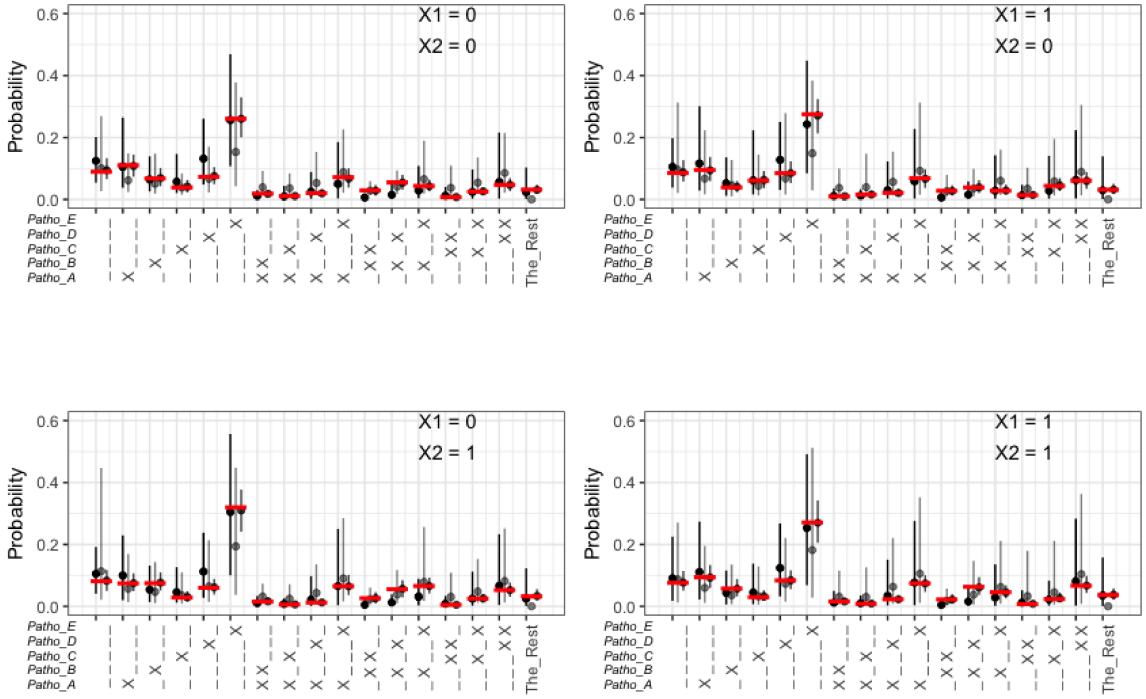

Fig. 3.5: **Singleton and doubleton etiology probability estimation comparison**: Each of the four plots in this figure stands for a specific stratum labeled by the top-right legend. In each plot, the x-axis includes 17 different etiological combinations. Each of the first 16 combinations from the left is denoted by a unique combination of '_' and 'X', where '_' means no infection and 'X' mean infection for the corresponding pathogen listed on the very left. The last combination is labeled by 'The_Rest' indicating the sum of all the rest possible combinations, e.g. tripletons, etc. For each combination, there are three vertical lines, of which the upper and lower bounds represent the 97.5th and 2.5th percentiles of the sampling distribution. The solid dots along these lines indicate the mean of the sampling distribution. The red horizontal lines mark the true values. From left to right, the three lines correspond to three scenarios: (left) LSC estimates in Study I (data has co-infection and low measurement quality) with SS measurements only available for pathogen D and E; (middle) pLCM-2 estimates in Study I with SS measurements only available for pathogen D and E; (right) LSC estimates in Study III (data has co-infection and high measurement quality).

$\widetilde{a}_k = 7.6$, $\widetilde{b}_k = 59.0$, $a_k^* = 6.0$, and $b_k^* = 1.3$ by range matching, and $a'_k = b'_k = 1$ for non-informativeness. Two categorical (binary) variables: age group and disease severity are taken into account, thus the etiology estimation is made for 4 strata. Regression coefficients including the interaction terms all have prior mean at zero. Other hyper-parameter settings are $\mu_\rho^* = -5, g_d = 6, h_d = 2, \sigma_\rho = \sigma_\beta = 2.2$.

Figure 3.6 shows the singleton and doubleton etiology probability estimates for Kenya site given by three models. As we can see, the singleton etiology probability estimates made by the LSC model are quite similar to the corresponding estimates made by pLCM-1. The doubleton estimates and 'The Rest' estimates made by the LSC model are mostly nearly zero, with an exception for the PNEU-HINF pair in the Age = 0 and Severity = 1 stratum. These two pathogens are both commensals of the human nasopharynx and have long been detected together in a multispecies biofilm in infected tissue Tikhomirova and Kidd (2013). A noticeable difference between pLCM-2 results and the other two models is that pLCM-2 attributes the etiology more to RHINO and RHINO-PNEU pair and less to pathogens other than these five. Singleton et al. Singleton et al. (2010) categorize respiratory viruses into two groups based on their contribution to disease. Group 1 includes viruses with a significantly greater contribution to respiratory symptoms, including RSV, metapneumovirus, certain para-influenza viruses, and influenza viruses. Group 2 viruses, including human rhinoviruses, adenoviruses, and coronaviruses, are less likely to be the single etiological pathogen of disease in children. Thus, it appears

the pLCM-2 model overestimates the contribution of singleton RHINO in the Age = 1 strata, and the final inference should be based on the LSC model. The etiology probabilities estimated by the LSC model are listed in table 3.7 and 3.8 where 'None Above' means infection by other pathogens that are not among the listed five candidates, and 'The Rest' means infection by any other possible combinations of the listed five candidates.

Table 3.7: Etiology probability estimates for Kenya site

	Age = 1, Severity = 0			Age = 0, Severity = 0		
	Mean	2.5%	97.5%	Mean	2.5%	97.5%
None_Above	0.3162	0.0303	0.5748	0.1351	0.0288	0.3003
RSV	0.264	0.1335	0.4435	0.5601	0.3669	0.7666
RHINO	0.2471	0.0397	0.6324	0.1335	0.0257	0.3566
HMPV_A_B	0.086	0.0212	0.1946	0.0898	0.0355	0.1805
PNEU	0.0411	0.0041	0.136	0.0304	0.0045	0.0881
HINF	0.0341	0.0013	0.1083	0.0304	0.0058	0.0893
RSV-RHINO	0.0015	0	0.0111	0.0025	0	0.0149
RSV-HMPV_A_B	0.0004	0	0.0022	0.0016	0	0.0081
RSV-PNEU	0.0002	0	0.0011	0.0005	0	0.0033
RSV-HINF	0.0014	0	0.0188	0.0041	0	0.0523
RHINO-HMPV_A_B	0.0011	0	0.0066	0.0014	0	0.0034
RHINO-PNEU	0.0013	0	0.0124	0.0016	0	0.0194
RHINO-HINF	0.0007	0	0.0083	0.0009	0	0.0123
HMPV_A_B-PNEU	0.0001	0	0.0005	0.0002	0	0.0006
HMPV_A_B-HINF	0.0002	0	0.0028	0.0005	0	0.0075
PNEU-HINF	0.0049	0	0.0236	0.0073	0	0.0268
The_Rest	0	0	0.0001	0.0001	0	0.0006

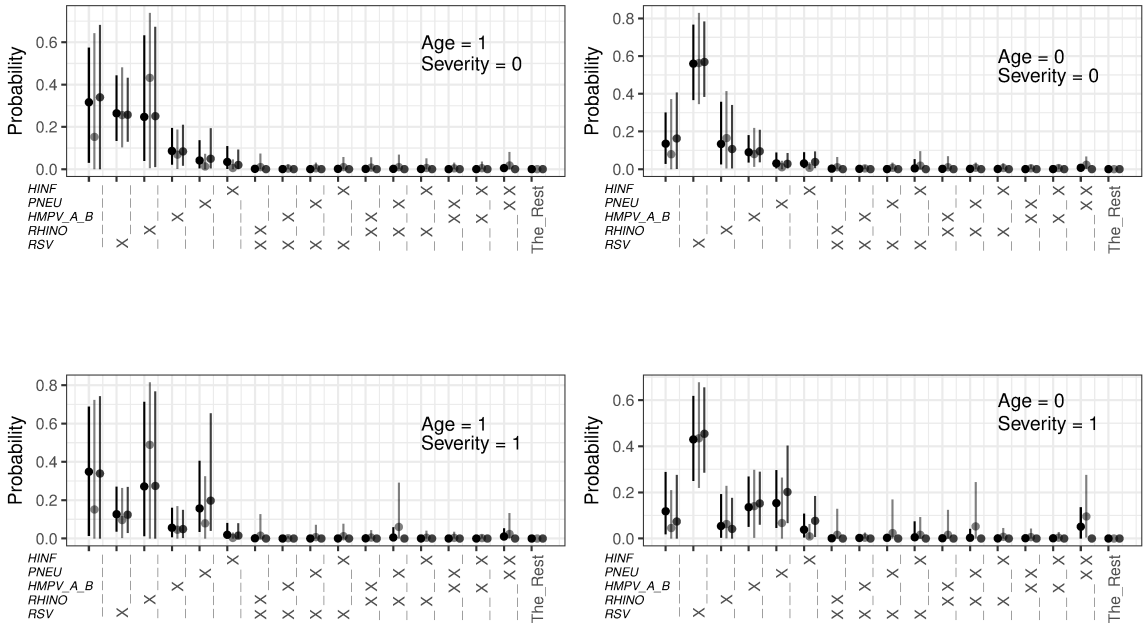

Fig. 3.6: **Singleton and doubleton etiology probability estimation for Kenya site**: The legends and labels in this figure have the same meaning as they are in figure 3.5, except that the pathogen names listed in the x-axis labels in this figure are the real pathogen abbreviations. The three vertical lines for each etiological combination correspond to the three models applied to the Kenya data set: (left) the LSC model; (middle) the pLCM-2 model; (right) the pLCM-1 model.

Table 3.8: Etiology probability estimates for Kenya site

	Age = 1, Severity = 1			Age = 0, Severity = 1		
	Mean	2.5%	97.5%	Mean	2.5%	97.5%
None_Above	0.3487	0.0143	0.6874	0.1181	0.0188	0.2886
RSV	0.1269	0.0365	0.2703	0.4295	0.2498	0.6179
RHINO	0.2718	0.0127	0.7124	0.0541	0.0034	0.1918
HMPV_A_B	0.0561	0.0066	0.16	0.1358	0.0512	0.268
PNEU	0.157	0.0346	0.4057	0.1535	0.0461	0.2964
HINF	0.0195	0.0003	0.0815	0.0383	0.0044	0.1073
RSV-RHINO	0.0012	0	0.0098	0.001	0	0.0067
RSV-HMPV_A_B	0.0001	0	0.0008	0.0025	0	0.0143
RSV-PNEU	0.0004	0	0.0028	0.003	0	0.0201
RSV-HINF	0.0005	0	0.0068	0.0054	0	0.0744
RHINO-HMPV_A_B	0.0009	0	0.0074	0.001	0	0.004
RHINO-PNEU	0.0055	0	0.0585	0.0031	0	0.0415
RHINO-HINF	0.0004	0	0.0053	0.0006	0	0.0073
HMPV_A_B-PNEU	0.0003	0	0.0013	0.0014	0	0.0063
HMPV_A_B-HINF	0.0001	0	0.0007	0.001	0	0.0118
PNEU-HINF	0.0105	0	0.053	0.0514	0.0001	0.1356
The_Rest	0	0	0.0002	0.0003	0	0.0026

3.6 Discussion and Future Work

In this work, we propose a new method, the latent sparse correlation (LSC) model, for pneumonia etiology estimation using non-Gold standard measurements. Under the partially-identifiable condition, this method finds a balance between the flexibility of latent variable representation and the model identifiability by proposing a sparse correlation structure of the latent multivariate binary variables. A MALA-within-Gibbs sampling algorithm is proposed for posterior approximation. Let K be the dimension of the latent state and let n be the sample size. By applying the pseudo-likelihood of latent variables, this MCMC algorithm takes $O(K^2 n)$ to finish

each iteration, which is scalable with respect to both K and n. Simulation studies show that this approach can provide estimation for the latent etiology distribution reasonably well while allowing arbitrary combinations of pathogen infection. In the PERCH data analysis, although we do not have the Gold standard measurements to validate our estimation, the results of the LSC model are consistent with published etiology research findings.

A limitation of this method is that its estimation accuracy relies on the following assumptions. (1) The measurements are conditionally independent given the true latent status. (2) The experts' knowledge used for setting the TPR priors does not contradict with the truth. (3) The correlation structure of the latent nodes is sign-consistent, that is, their correlations are either all non-negative or all non-positive. Future work could be used to relax the above assumptions and make this method more generally applicable. For example, we could borrow the nested structure proposed in Wu et al. (2017) to model the conditional dependence among measurements. We can modify the D matrix in the LSC model by adding a third state to allow for synergic effects between pathogens, but this could compromise the model identifiability. In addition, although the proposed MCMC algorithm is fairly scalable, when K gets moderately large, say 20, this algorithm could take hours to run. In practice, when researchers need to fit the model with a few prototype specifications or test for the impact of different prior settings, the computation burden is still too much given limited resources. Therefore, additional works can be done to

develop a faster estimating procedure without significant sacrifice in estimation accuracy. Popular alternatives to MCMC for posterior approximation include approximate Bayesian computation (ABC) techniques (Sunnåker et al., 2013), integrated nested Laplace approximation (INLA) (Rue et al., 2009), variational Bayesian inference (VBI) (Blei et al., 2006), and the closely related expectation-propagation (EP) (Minka, 2001).

Chapter 4

Fast Variational Inference of the Latent Sparse Correlation Model for Etiology Estimation

Abstract

Pneumonia, infection of the lung, is the number one cause of death for children under five. It can be caused by more than 30 different pathogens. The Pneumonia Etiology Research for Childhood Pneumonia (PERCH) study is a multi-country case-control study to estimate the frequency with which each pathogen causes pneumonia (etiology distribution). This goal is challenging because sampling directly from a child's lung (gold standard) is not typically feasible. Rather pathogens are enumerated by PCR from multiple peripheral sites including the nose and blood. These measurements are of varying quality with imperfect sensitivity and specificity. It is shown that the Bayesian Latent Sparse Correlation Model (BLSCM) with MCMC estimation provides accurate approximation to the latent etiology distribution in simulation studies. But the MCMC algorithm can get very slow with moderately large number of latent variables. In this work, we propose a fast variational inference algorithm for approximating the posterior of the BLSCM. Simulation studies show that this approach can provide reasonably accurate estimation for the etiology distribution using much shorter time. Its application to the PERCH data set provides etiology estimate that is consistent with published etiology research findings and gives insight into the possible coinfection patterns in childhood pneumonia patients.

4.1 Introduction

Pneumonia is a form of acute respiratory infection of the lungs. A variety of pathogens can cause the infection, including bacteria, viruses, mycobacteria and fungi (Hirama et al., 2011). Although the majority of child pneumonia cases are non-severe and can be managed in local primary health care facilities, the severe/very severe cases may result in death, especially in developing countries (Levine et al., 2012). In fact, pneumonia is the single largest infectious cause of mortality among children under five years of age, with an estimate of 0.92 million deaths per year accounting for 16% of the total 5.9 million childhood deaths worldwide (Liu et al., 2016; Black et al., 2010). It has been over thirty years since the last comprehensive study of pneumonia etiology (Shann, 1986; World Health Organization et al., 1990). Significant changes have taken place in vaccine use, HIV infection, living conditions, nutrition, and access to health care (Levine et al., 2012). These changes will certainly modify the distribution of pathogens, the transmission, and the natural history of infection, which will make the understandings of pneumonia etiology based on the early studies invalid. The Pneumonia Etiology Research for Childhood Pneumonia (PERCH) project is a multi-country case-control study to estimate the frequency with which each pathogen causes pneumonia (etiology distribution) (Levine et al., 2012). This goal is challenging because sampling directly from a childs lung is not typically feasible given the invasiveness of the procedure

(Levine et al., 2012; Hammitt et al., 2012). The actual pathogen(s) that infect the lung, therefore, can only be inferred from multiple peripheral measurements (non-gold standard data) with imperfect sensitivity and specificity (Murdoch et al., 2012; Hammitt et al., 2012).

Conventional approaches to etiology estimation are rule-based without incorporating evidence from all sources. Previous statistical methods (Wu et al., 2015, 2017) only allow single-pathogen cause or pre-defined sets of joint cause. To allow for arbitrary combinations of pathogens infecting the lung, we proposed a Bayesian latent sparse correlation model (BLSCM) in Chapter 3 of this thesis. In BLSCM, a multivariate binary vector L is defined to be latent variable that denotes the actual status of the lung. Given the latent variable L, the measurements for different pathogens from various peripheral sites are assumed to be conditionally independent and parameterized by their corresponding true positive rates (TPR) and false positive rates (FPR). Also, informative priors are used for the TPRs. The priors are specified as independent Beta distributions where the hyper-parameters that are determined by a credible interval matching procedure. A MALA-within-Gibbs sampling algorithm with pseudo-likelihood is developed for posterior approximation. Let K be the dimension of L and let n be the number of case observations. By applying the pseudo-likelihood of latent variables, this MCMC algorithm takes $O(K^2n)$ to finish each iteration, which is fairly scalable with respect to both K and n. However, when K gets moderately large, say 20, this algorithms could take hours to run. In

practice, when researchers need to fit the model with a few prototype specifications or test for the impact of different prior settings, the computation burden is still too much given limited resources. Therefore, a much faster estimating procedure without significant sacrifice in estimation accuracy is desired.

Other than MCMC, the most widely used methods for approximating the posterior distribution of latent structure models include approximate Bayesian computation (ABC) techniques (Sunnåker et al., 2013), integrated nested Laplace approximation (INLA) (Rue et al., 2009), variational Bayesian inference (VBI) (Blei et al., 2006), and the closely related expectation-propagation (EP) (Minka, 2001). All but INLA, which requires continuous latent variable, can be adapted to approximate the posterior of the Bayesian latent sparse correlation model. In a typical ABC algorithm, each iteration starts with proposing a candidate of parameter value, then a set of data is simulated based on the proposed parameter, finally the proposal is accepted if the distance between the simulated data and the observed data is smaller than a threshold. This approach is also sampling-based. It is usually faster than MCMC when simulating data based on the model can be accomplished more efficiently than evaluating the likelihood function, but it is not the case in BLSCM with pseudo-likelihood. Thus, we may not get much speed-up by using ABC techniques. VBI and EP are both popular deterministic approximation methods for Bayesian latent variable models. They both restrict the approximation in a simpler family of distributions and seek to find the closest member in that family to

the true posterior. The difference between them lies in the metric of "closeness". VBI optimizes the evidence lower bound (ELBO) which guarantees convergence to local optimum, but EP does not have such guarantee and could converge to a saddle point. In this work, we propose a variational Bayesian inference algorithm for approximating the posterior of the BLSCM.

4.2 Bayesian Latent Sparse Correlation Model

4.2.1 Likelihood Function

For any individual latent state, we call the binary measurements with perfect sensitivity and specificity Gold Standard (GS) measurements. The type of measurements with perfect specificity, but imperfect sensitivity are called Silver Standard (SS) measurements, and the measurements with both imperfect sensitivity and specificity are Bronze Standard (BS) measurements. For each independent observation i, let X_i be the covariate vector, including age, gender, etc., let Y_i be the group assignment indicator depending on the study design, for example: case ($Y_i = 1$) vs. control ($Y_i = 0$). Let $L_i = (l_{i1}, l_{i2}, \ldots, l_{iK})$ be the K-dimensional multivariate binary latent variable, then the list of measurements can be described by three K-dimensional binary vectors: M_i^{GS} (if available), M_i^{SS}, and M_i^{BS}. Let $\gamma \in [0,1]^K$ and $\delta \in [0,1]^K$ represent the True Positive Rate (TPR) and False Posi-

tive Rate (FPR) for BS measurements respectively, and let $\eta \in [0,1]^K$ be the TPR for SS measurements. Under the assumption that SS and BS measurements are conditionally independent given the latent states, we have the conditional distribution of the measurements given the latent states.

For the BS measurements:

$$
\begin{aligned}
& P(M_i^{BS}|Y_i, L_i; \gamma, \delta,) \qquad (4.1)\\
=& \prod_{k=1}^{K} P(M_{ik}^{BS}|Y_i, L_i; \gamma, \delta,) \\
=& \exp\Big\{ \sum_{k=1}^{K} \Big[\log(1-\delta_k) + M_{ik}^{BS} \log \frac{\delta_k}{1-\delta_k} + Y_i L_{ik} \log \frac{1-\gamma_k}{1-\delta_k} + Y_i L_{ik} M_{ik}^{BS} \log \frac{\gamma_k(1-\delta_k)}{\delta_k(1-\gamma_k)} \Big] \Big\}.
\end{aligned}
$$

For the SS measurements:

$$
\begin{aligned}
& P(M_i^{SS}|Y_i, L_i; \eta) \qquad (4.2)\\
=& \exp\Big\{ \sum_{k=1}^{K} L_{ik} M_{ik}^{SS} \log \frac{\eta_k}{1-\eta_k} + L_{ik} \log(1-\eta_k) + M_{ik}^{SS} \log L_{ik} \Big\}.
\end{aligned}
$$

The latent variable likelihood with sparse correlation:

$$
\begin{aligned}
& P(L_i|X_i; \beta, \rho, D) \qquad (4.3)\\
=& \exp\Big\{ \sum_{k=1}^{K} (L_{ik} X_i^T \beta_k + \rho \sum_{k \neq k'} L_{ik} L_{ik'} D_{kk'}) - A(\beta, \rho, D) \Big\},
\end{aligned}
$$

where

$$A(\beta, \rho, D) = \log \Big\{ \sum_{\boldsymbol{l} \in \{0,1\}^K} \exp \big[\sum_{k=1}^{K} (l_k X_i^T \beta_k + \rho \sum_{k' \neq k} l_k l_{k'} D_{kk'}) \big] \Big\}, \tag{4.4}$$

and D is a symmetric random matrix, such that $\forall k_1, k_2 \in \{1, 2, \ldots, K\}, k_1 < k_2$,

$$D_{k_1 k_2} \sim \text{Bernolli}(d_{k_1 k_2}), D_{k_2 k_1} = D_{k_1 k_2}$$

$$d_{k_1 k_2} \sim \text{Beta}(g_d, h_d) \text{ , .}$$

Then the true augmented likelihood function is the product of term 4.1, 4.2 and 4.3. Note that the normalizing term $A(\beta, \rho, D)$ requires a summation over all 2^K categories, thus the true likelihood in equation 4.3 is replaced with the following pseudo-likelihood (Besag, 1974).

$$\begin{aligned} &\text{pL}(L_{i1}, L_{i2}, \ldots, L_{iK} | X_i, \boldsymbol{\beta}, \rho, \boldsymbol{D}) \\ &= \exp \Big\{ \sum_{k=1}^{K} \big[L_{ik} X_i^T \beta_k - \log \big(1 + \exp(X_i^T \beta_k + \rho \sum_{k' \neq k} L_{ik'} D_{kk'}) \big) \big] \Big\} \end{aligned} \tag{4.5}$$

4.2.2 Prior Specification

According to the likelihood function, the parameters of interest are $(\beta, \rho, D, \eta, \gamma, \delta)$. In the joint prior distribution, we assume these parameters are independent. With the control group data available, we can use a Beta$(a'_k = 1, b'_k = 1)$ distribution as

the non-informative prior of each δ_k. But for η, γ, informative priors are required. We adopt the same configuration proposed in pLCM Wu et al. (2015), where the joint prior distribution is specified as a product of the individual Beta priors. Let $(\widetilde{a}_k, \widetilde{b}_k)$ be the hyper-parameter that defines the prior of η_k, and let (a_k^*, b_k^*) be the hyper-parameter for γ_k, then the value of $\widetilde{a}_k, \widetilde{b}_k, a_k^*, b_k^*$ are calculated by matching the quantiles of the Beta distributions with the experts' best knowledge on the TPRs and FPRs of each type of measurement. For example, in the Pneumonia etiology research application, the scientists may believe there is 95 $\%$ of chance that the TPR of a SS measurement (e.g. blood culture) is between 0.01 and 0.2, then the value of $(\widetilde{a}_k, \widetilde{b}_k)$ is determined by setting the 2.5th and 97.5th percentile of Beta$(\widetilde{a}_k, \widetilde{b}_k)$ to 0.01 and 0.2 respectively and solving the equation.

For β, the prior is specified as a product of the individual Gaussian distributions, with mean $\mu_{\beta_k}^* = 0$, and variance $\sigma_{\beta_k}^{*2}$. In most cases, we recommend using a relatively large value for the prior variance as we want to stay non-informative for these set of parameters unless strong prior knowledge is present. The prior of ρ is also assumed as a Gaussian distribution, with mean μ_ρ^*, and variance σ_ρ^{*2}, but it requires some extent of prior knowledge to set the values of these two hyper-parameters. Essentially, the sign of μ_ρ^* defines whether the latent nodes are competing with each other or promoting each other. As defined in formula 3.1, $D_{kk'}$ has a Bernoulli prior with parameter $d_{kk'}$, and each $d_{kk'}$ is a Beta prior that shares the same set of hyper-parameters g_d and h_d.

4.3 Variational Inference of the Latent Sparse Correlation Model

Variational Bayesian inference Xing et al. (2002); Wainwright et al. (2008) is a class of method that approximates the posterior distribution $P(\cdot)$ by a model $Q(\cdot)$ that has some tractable form, such as the exponential family, and the Kullback-Leibler (KL) divergence between Q and P is minimized. In this work, we develop our approximation algorithm based on the mean-field variational family, which assumes that Q is completely factorizable with respect to each parameter component. Let $\Lambda = (\beta, \rho, D, \eta, \gamma, \delta, L)$, and let $\pi(\Lambda|M, X)$ be the posterior distribution of interest. We are interested in approximating $\pi(\Lambda|M, X)$ by $Q(\Lambda) = \prod_{\lambda \in \Lambda} q(\lambda)$, which minimizes the KL divergence between $\pi(\Lambda|M, X)$ and $Q(\Lambda)$. This is also equivalent to maximizing the evidence lower bound (ELBO):

$$Q(\Lambda) = \text{argmax}\Big\{\mathbb{E}_Q\Big[\log \frac{\pi(\Lambda, M, X)}{Q(\Lambda)}\Big]\Big\}. \tag{4.6}$$

It is shown (Bishop, 2006) that the optimal variational factors $q^*(\cdot)$ is proportional to the exponentiated expected log of its conditional density given all the other variables, that is,

$$q^*(\lambda) \propto \exp\Big\{\mathbb{E}_{Q(-\lambda)}\big[\log \pi(\Lambda|M, X)\big]\Big\}. \tag{4.7}$$

Based on the mean-field family assumption that the variational factors are mutually independent, the right hand side of equation 4.7 does not involve $q(\lambda)$. Therefore, formula 4.7 provides a valid coordinate update function for λ that underlies a coordinate ascent algorithm. Next, we show the exact or approximated form of the update functions for $(\beta, \rho, D, \eta, \gamma, \delta, L)$.

- **q(η), q(γ), q(δ)**

Let $q_{l_{ik}} = \mathbb{E}_Q(L_{ik}|X_i)$. We can show that for the TPRs and FPRs, the approximated posterior distributions are Beta distributions, that is, $q(\eta_k) = \text{Beta}(\widetilde{A}_k, \widetilde{B}_k)$, $q(\gamma_k) = \text{Beta}(A_k^*, B_k^*)$, and $q(\delta_k) = \text{Beta}(A_k', B_k')$, where we assume the first n_{case} observations are from the case group, n is the total sample size, and

$$
\begin{aligned}
\widetilde{A}_k &= \sum_{i=1}^{n_{case}} q_{l_{ik}} M_{ik}^{SS} + \widetilde{a}_k \\
\widetilde{B}_k &= \sum_{i=1}^{n_{case}} q_{l_{ik}} (1 - M_{ik}^{SS}) + \widetilde{b}_k \\
A_k^* &= \sum_{i=1}^{n} Y_i q_{l_{ik}} M_{ik}^{BS} + a_k^* \\
B_k^* &= \sum_{i=1}^{n} Y_i q_{l_{ik}} (1 - M_{ik}^{BS}) + b_k^* \\
A_k' &= \sum_{i=1}^{n} M_{ik}^{BS} (1 - Y_i q_{l_{ik}}) + a_K' \\
B_k' &= n + \sum_{i=1}^{n} (Y_i q_{l_{ik}} M_{ik}^{BS} - Y_i q_{l_{ik}} - M_{ik}^{BS}) + b_K'.
\end{aligned}
\tag{4.8}
$$

- **q(L_{ik})**

The approximated posterior distribution of the latent variable $q(L_{ik})$ is a Bernoulli distritbution with prameter $q_{l_{ik}}$. Let $\psi(\cdot)$ denote the Digamma function, and let $q_{D_{kk'}} = \mathbb{E}_Q(D_{kk'})$, $\mu_\rho = \mathbb{E}_Q(\rho)$, and $\mu_{\theta_{ik}} = X_i^T \mathbb{E}_Q(\beta_k)$, then we can show that

$$q_{l_{ik}} = \frac{\exp(H_{ik} + \mu_\rho \sum_{k' \neq k} q_{l_{ik'}} q_{D_{kk'}})}{1 + \exp(H_{ik} + \mu_\rho \sum_{k' \neq k} q_{l_{ik'}} q_{D_{kk'}})}, \tag{4.9}$$

where for the latent variable L_{ik}:

1. When SS measurement is available and $M_i^{SS} k = 1$, $q_{l_{ik}} = 1$.

2. When SS measurement is availabe and $M_i^{SS} k = 0$, but BS measurement is not available, then

$$H_{ik} = \psi(\widetilde{B}_k) - \psi(\widetilde{A}_k + \widetilde{B}_k).$$

3. When only BS measurement is available, then

$$\begin{aligned} H_{ik} =& M_{ik}^{BS}[\psi(A_k^*) - \psi(A_k')] + (1 - M_{ik}^{BS})[\psi(B_k^*) - \psi(B_k')] \\ &+ \mu_{\theta_{ik}} - \psi(A_k^* + B_k^*) + \psi(A_k' + B_k') \\ =& H_{ik}^{BS}. \end{aligned}$$

4. When SS and BS measurements are both availabe and $M_i^{SS}k = 0$, then

$$H_{ik} = H_{ik}^{BS} + \psi(\widetilde{B}_k) - \psi(\widetilde{A}_k + \widetilde{B}_k).$$

- **q(β_k)**

Suppose the design matrix of the covariates has m columns, then $X_i = (x_{i1}, \ldots, x_{im})^T$, $\beta_k = (\beta_{1k}, \ldots, \beta_{mk})^T$, and $\theta_{ik} = X_i^T \beta_k$. By using pseudo-likelihood function to approximate the normalizing constant, we have

$$\begin{aligned}
&\exp\Big\{Y_i \mathbb{E}_Q\big[\sum_{k=1}^{K} L_{ik}\theta_{ik} - A(\theta, \rho, D)\big]\Big\} \\
\approx &\exp\Big\{Y_i \sum_{k=1}^{K}\big[q_{l_{ik}}\theta_{ik} - \mathbb{E}_Q \log\big(1 + \exp(\theta_{ik} + \rho \sum_{k' \neq k} L_{ik'} D_{kk'})\big)\big]\Big\}.
\end{aligned}$$

Let $R_{ik} = \sum_{k' \neq k} L_{ik'} D_{kk'}$, we can show that

$$\mathbb{V}ar(\rho R_{ik}) = (\sigma_\rho^2 + \mu_\rho^2) \sum_{k' \neq k} q_{l_{ik'}} q_{D_{kk'}} (1 - q_{l_{ik'}} q_{D_{kk'}}) + \sigma_\rho^2 [\sum_{k' \neq k} q_{l_{ik'}} q_{D_{kk'}}]^2.$$

By applying Laplacian Approximation, we can show that

$$\begin{aligned}
q(\beta_{jk}) \approx \exp\Big\{ \sum_{i=1}^{n} Y_i \Big[q_{l_{ik}} x_{ij} \beta_{jk} - \log(1 + e^*) - \frac{e^*}{2(1 + e^*)^2}\big(\mathbb{V}ar(\rho R_{ik}) + \sum_{j' \neq j} x_{ij'}^2 \sigma_{\beta_{j'k}}^2\big)\Big] \\
- \frac{1}{2}(\beta_{jk} - \mu_{\beta_{jk}}^*)^2 / \sigma_{\beta_{jk}}^{*2} \Big\}, \qquad (4.10)
\end{aligned}$$

where

$$e^* = \exp\Big\{x_{ij}\beta_{jk} + X_{i(-j)}^T \mu_{\beta_{(-j)k}} + \mu_\rho \sum_{k'\neq k} q_{l_{ik'}} q_{D_{kk'}}\Big\}.$$

Then $\mu_{\beta_{jk}} \approx$ argmax$q(\beta_{jk})$ and $\sigma^2_{\beta_{jk}} \approx -\Big[\frac{\partial^2 \log q(\beta_{jk})}{\partial \beta_{jk}^2}\Big|_{\beta_{jk}=\mu_{\beta_{jk}}}\Big]^{-1}$.

- **q(ρ)**

With pseudo-likelihood approximation, we have

$$\begin{aligned}
&\exp\Big\{Y_i \mathbb{E}_Q\big[\rho \sum_{k=1}^{K}\sum_{k'\neq k} L_{ik}L_{ik'}D_{kk'} - A(\beta,\rho,D)\big]\Big\} \\
\approx &\exp\Big\{Y_i \sum_{k=1}^{K}\big[\rho \sum_{k'\neq k} q_{l_{ik}} q_{l_{ik'}} q_{D_{kk'}} - \mathbb{E}_Q \log\big(1+\exp(x_i^T\beta_k + \rho\sum_{k'\neq k} L_{k'}D_{kk'})\big)\big]\Big\}.
\end{aligned} \tag{4.11}$$

By Le Cam's theorem, the distribution of R_{ik} can be approximated by Poisson$(\sum_{k'\neq k} q_{l_{ik'}} q_{D_{kk'}})$, thus

$$\begin{aligned}
q(\rho) \propto \exp\Big\{ \sum_{i=1}^{n} Y_i \sum_{k=1}^{K} \big[\rho \sum_{k'\neq k} q_{l_{ik}} q_{l_{ik'}} q_{D_{kk'}} \\
- \sum_{j=0}^{K-1} \text{Poi}(j;\lambda = \sum_{k'\neq k} q_{l_{ik'}} q_{D_{kk'}})\Big(\log(1+e^{X_i^T\mu_{\beta_k}+j\rho}) \\
+ \frac{e^{X_i^T\mu_{\beta_k}+j\rho}}{2(1+e^{X_i^T\mu_{\beta_k}+j\rho})^2}\mathbb{V}ar(X_i^T\beta_k)\Big)\big]\Big\}(\rho-\mu_\rho^*)^2/\sigma_\rho^{*2}.
\end{aligned} \tag{4.12}$$

Then $\mu_\rho \approx$ argmax $q(\rho)$, and $\sigma^2_\rho \approx -\Big[\frac{\partial^2 \log q(\rho)}{\partial \rho^2}\Big|_{\rho=\mu_\rho}\Big]^{-1}$.

- **q(D)**

Based on formula (10), we can show that for all k_1 and $k_2 \in \{1, 2, \dots K\}$ and $k_1 \neq k_2$,

$$
\begin{aligned}
q(D_{k_1k_2}) \propto \exp\Big\{ \sum_{i=1}^{n} Y_i \Big[& 2\mu_\rho q_{l_{ik_1}} q_{l_{ik_2}} D_{k_1k_2} \\
& - \mathbb{E}_Q \log(1 + e^{X_i^T \beta_{k_1} + \rho \sum_{k\neq k_1, k\neq k_2} L_{ik} D_{kk_1} + \rho L_{ik_2} D_{k_1k_2}}) \\
& - \mathbb{E}_Q \log(1 + e^{X_i^T \beta_{k_2} + \rho \sum_{k\neq k_1, k\neq k_2} L_{ik} D_{kk_2} + \rho L_{ik_1} D_{k_1k_2}}) \Big] \\
& + D_{k_1k_2} \mathbb{E}_Q (\log \frac{d_{k_1k_2}}{1 - d_{k_1k_2}}) \Big\}. \qquad (4.13)
\end{aligned}
$$

Let $S^{(i)}_{k_1k_2} = X_i^T \beta_{k_1} + \rho \sum_{k\neq k_1, k\neq k_2} L_{ik} D_{kk_1}$, $T^{(i)}_{k_1k_2} = \mathbb{E}_Q \log(1 + e^{S^{(i)}_{k_1k_2} + \rho L_{ik_2}})$, $T'^{(i)}_{k_1k_2} = \mathbb{E}_Q \log(1 + e^{S^{(i)}_{k_1k_2}})$, Let $A_{d_{k_1k_2}}$ and $B_{d_{k_1k_2}}$ denote the Beta posterior parameters of $d_{k_1k_2}$, then for all $k_1 < k_2$, we have

$$
\begin{aligned}
T^{(i)}_{k_1k_2} \approx & \log(1 + e^{X_i^T \mu_{\beta_{k_1}} + \mu_\rho \sum_{k\neq k_1, k\neq k_2} q_{l_{ik}} q_{D_{kk_1}} + \mu_\rho q_{l_{ik_2}}}) \\
& + \frac{e^{X_i^T \mu_{\beta_{k_1}} + \mu_\rho \sum_{k\neq k_1, k\neq k_2} q_{l_{ik}} q_{D_{kk_1}} + \mu_\rho q_{l_{ik_2}}}}{2(1 + e^{X_i^T \mu_{\beta_{k_1}} + \mu_\rho \sum_{k\neq k_1, k\neq k_2} q_{l_{ik}} q_{D_{kk_1}} + \mu_\rho q_{l_{ik_2}}})^2} [\mathbb{V}ar(S^{(i)}_{k_1k_2}) + \mathbb{V}ar(\rho L_{ik_2})], \\
T'^{(i)}_{k_1k_2} \approx & \log(1 + e^{X_i^T \mu_{\beta_{k_1}} + \mu_\rho \sum_{k\neq k_1, k\neq k_2} q_{l_{ik}} q_{D_{kk_1}}}) \\
& + \frac{e^{X_i^T \mu_{\beta_{k_1}} + \mu_\rho \sum_{k\neq k_1, k\neq k_2} q_{l_{ik}} q_{D_{kk_1}}}}{2(1 + e^{X_i^T \mu_{\beta_{k_1}} + \mu_\rho \sum_{k\neq k_1, k\neq k_2} q_{l_{ik}} q_{D_{kk_1}}})^2} \mathbb{V}ar(S^{(i)}_{k_1k_2}),
\end{aligned}
$$

and

$$q_{D_{k_1k_2}} = \mathbb{E}_Q(D_{k_1k_2}) \tag{4.14}$$
$$= \frac{\exp\left\{\sum_{i=1}^n Y_i[2\mu_\rho q_{l_{ik_1}} q_{l_{ik_2}} - T^{(i)}_{k_1k_2} - T^{(i)}_{k_2k_1}] + \mathbb{E}_Q(\log \frac{d_{k_1k_2}}{1-d_{k_1k_2}})\right\}}{\exp\left\{\sum_{i=1}^n Y_i[2\mu_\rho q_{l_{ik_1}} q_{l_{ik_2}} - T^{(i)}_{k_1k_2} - T^{(i)}_{k_2k_1}] + \mathbb{E}_Q(\log \frac{d_{k_1k_2}}{1-d_{k_1k_2}})\right\} + \exp\left\{-\sum_{i=1}^n Y_i[T'^{(i)}_{k_1k_2} + T'^{(i)}_{k_2k_1}]\right\}}$$

where

$$\begin{aligned}
\mathbb{E}_Q(\log \frac{d_{k_1k_2}}{1-d_{k_1k_2}}) =& \psi(A_{d_{k_1k_2}}) - \psi(B_{d_{k_1k_2}}) \\
\mathbb{V}ar(S^{(i)}_{k_1k_2}) =& \mathbb{V}ar(X_i^T\beta_{k_1}) + (\sigma_\rho^2 + \mu_\rho^2) \sum_{k\neq k_1, k\neq k_2} q_{l_{ik}} q_{D_{kk_1}} (1 - q_{l_{ik}} q_{D_{kk_1}}) \\
& + \sigma_\rho^2 [\sum_{k\neq k_1, k\neq k_2} q_{l_{ik}} q_{D_{kk_1}}]^2 \\
\mathbb{V}ar(\rho L_{ik}) =& (\sigma_\rho^2 + \mu_\rho^2) q_{l_{ik}} - \mu_\rho^2 q_{l_{ik}}^2.
\end{aligned}$$

- **q(**$d_{k_1k_2}$**)**

For any $k_1 < k_2$, we show that the approximated posterior distribution of $d_{k_1k_2}$ is a Beta($A_{d_{k_1k_2}}, B_{d_{k_1k_2}}$):

$$q(d_{k_1k_2}) \propto \exp\Big\{(q_{D_{k_1k_2}} + g_d - 1)\log d_{k_1k_2} + (1 - q_{D_{k_1k_2}} + h_d - 1)\log(1 - d_{k_1k_2})\Big\}, \tag{4.15}$$

where $A_{d_{k_1k_2}} = q_{D_{k_1k_2}} + g_d$ and $B_{d_{k_1k_2}} = 1 - q_{D_{k_1k_2}} + h_d$.

Algorithm 1 Coordinate Ascent Variational Inference algorithm

1: Initialize all parameters (Λ) randomly.
2: Update $\widetilde{A}, \widetilde{B}, A^*, B^*, A', B'$ with equation 4.8.
3: **for** each i, k **do**
4: Update $q_{l_{ik}}$ with equation 4.9.
5: **end for**
6: **for** each j, k **do**
7: Optimize for $\mu_{\beta_{jk}} = \text{argmax} q(\beta_{jk})$
8: Calculate $\sigma^2_{\beta_{jk}} = -\left[\frac{\partial^2 \log q(\beta_{jk})}{\partial \beta^2_{jk}}\Big|_{\beta_{jk}=\mu_{\beta_{jk}}}\right]^{-1}$
9: **end for**
10: Calculate $\mu_\rho = \text{argmax } q(\rho)$, and $\sigma^2_\rho = -\left[\frac{\partial^2 \log q(\rho)}{\partial \rho^2}\Big|_{\rho=\mu_\rho}\right]^{-1}$
11: **for** each $k_1 < k_2$ **do**
12: Update $q_{D_{k_1 k_2}}$ with equation 4.14.
13: Update $q_{D_{k_2 k_1}} \leftarrow q_{D_{k_1 k_2}}$.
14: **end for**
15: **for** each $k_1 < k_2$ **do**
16: Update $A_{d_{k_1 k_2}}, B_{d_{k_1 k_2}}$ with equation 4.15.
17: Update $A_{d_{k_2 k_1}} \leftarrow A_{d_{k_1 k_2}}$, and $B_{d_{k_2 k_1}} \leftarrow B_{d_{k_1 k_2}}$.
18: **end for**
19: **if** Convergence is reached **then return** Λ
20: **else**
21: **Go to** step 2.
22: **end if**

Using the update functions of these variational factors, we construct the Coordinate Ascent Variational Inference (CAVI) algorithm (Bishop, 2006) described in Algorithm chart 1. As the ELBO is not always a convex function, the CAVI algorithm only guarantees convergence to a local optimum. Thus it is recommended to initiate the algorithm multiple times from random values then pick the best fit. The results of CAVI include the approximate posterior means of the parameters and latent variables. Then we can use these point estimates to recover the latent distribution and the likelihood value of the observed data. When K is small, the latent

distribution function can be calculated exactly using formula 4.3 where β, ρ, D are replaced by point estimates μ_β, μ_ρ, q_D. The transition probability matrix from L to M can also be calculated exactly. Thus we can integrate out L from the augmented likelihood and get the observed likelihood, with which we can compute DIC for prior selection. Namely, we pick the prior combination that produces the smallest DIC. However, as K gets large, the time complexity of the above calculation grows exponentially, therefore we propose to use the following pseudo-likelihood function to approximate the latent distribution.

$$\begin{aligned} & P(L_{new} = l | X_{new}; \mu_\beta, \mu_\rho, q_D) \\ \approx & \exp\Big\{ \sum_{k=1}^{K} \big[l_k X_{new}^T \mu_{\beta_k} - \log\big(1 + \exp(X_{new}^T \mu_{\beta_k} + \mu_\rho \sum_{k' \neq k} l_{k'} q_{D_{kk'}})\big)\big] \Big\} \end{aligned}$$

But even with the pseudo-likelihood approximation of the latent distribution, integrating out L is still intractable with large K. We propose a Monte Carlo approximation to the observed likelihood. First, simulate L based on the pseudo-likelihood approximated distribution function. Second, simulate the measurement data given L and the point estimates of δ, γ, η. For a large enough sample size, we can estimate the distribution function of M as a vector of multinomial probabilities with good accuracy. Last, compute the likelihood value and DIC based on the above Monte Carlo estimate.

4.4 Simulation Study

4.4.1 Design of Studies

We conducted a set of simulation studies to empirically evaluate the effectiveness of the LSC model under different situations. Specifically, we measure the accuracy of the recovered latent distribution, and how the accuracy changes with different measurement qualities and different prior specifications. For synthetic data generation, let there be five latent nodes $(A, B, \ldots, E)$, generated from the same multivariate binary distribution, in which the nodes are mostly negatively correlated. This true latent distribution is set as a regular QE model, where $\Theta_1 = (-1.5, -1.0, -0.5, 0.5, 1.0)$ and the association parameters are summarized in table 4.1. As we can see, node A and B are conditionally independent, node E is strongly negatively correlated with other nodes, and the negative correlation between A/C, B/D, and C/D are weaker than other pairs. Given the latent nodes L, the BS and SS measurements are then generated with two levels of qualities. Shown in table 4.2, at the high-quality level, $\text{TPR}^{(SS)} \approx 0.8, \text{TPR}^{(BS)} \approx 0.8, \text{FPR}^{(BS)} \approx 0.2$; at the low-quality level, $\text{TPR}^{(SS)} \approx 0.5, \text{TPR}^{(BS)} \approx 0.7, \text{FPR}^{(BS)} \approx 0.4$. As a result, for each observation, we have 10 (5 BS + 5 SS) high-quality measurements and 10 low-quality measurements.

Informative priors for η, γ, ρ, D are used for model identification. For η, γ, we adopt the percentile matching method proposed in pLCM Wu et al. (2015) to set

Table 4.1: Association Parameters

	A	B	C	D	E
A					
B	0				
C	-1	-2			
D	-2	-1	-1		
E	-2	-2	-2	-2	

Table 4.2: Measurement Quality Parameters

Quality	Parameter	A	B	C	D	E
High	TPR$^{(SS)}$	0.80	0.90	0.70	0.85	0.75
	TPR$^{(BS)}$	0.90	0.70	0.80	0.80	0.75
	FPR$^{(BS)}$	0.25	0.10	0.15	0.20	0.15
Low	TPR$^{(SS)}$	0.50	0.60	0.40	0.55	0.45
	TPR$^{(BS)}$	0.80	0.60	0.70	0.70	0.65
	FPR$^{(BS)}$	0.45	0.30	0.35	0.40	0.35

values for $\widetilde{a}_k, \widetilde{b}_k, a_k^*, b_k^*$. In the simulation study, we assume that our the expert knowledge on the TPRs are reasonable. For example, when studying the model performance with high-quality measurements, we know the true values of TPR$^{(SS)}$ and TPR$^{(BS)}$, thus our reasonable prior knowledge is that there is 95 $\%$ of chance that the TPR of a SS(or BS) measurement is between 0.6 and 0.99. Accordingly, we set $\widetilde{a}_k = a_k^* = 8.62$ and $\widetilde{b}_k = b_k^* = 1.41$. When studying model performance with low-quality measurements, our prior knowledge also changes. The intervals for TPR$^{(SS)}$ becomes $(0.3, 0.7)$ and the intervals for TPR$^{(BS)}$ becomes $(.5, 0.9)$. Therefore, we set $\widetilde{a}_k = \widetilde{b}_k = 11.26$, $a_k^* = 12.7$ and $b_k^* = 4.8$. For ρ and D, it is difficult to convert the experts' knowledge directly to the values of hyper-parameters, so we need to determine the values of $\mu_\rho^*, \tau_\rho^*, g_d, h_d$ through a grid search. The grid is defined by the unique combinations of $\mu_\rho^* \in \{-1, -3.5, -7\}, \tau_\rho^* \in \{1, 5, 10, 15\}, g_d \in \{1, 4, 8, 32\}$,

and $h_d \in \{1, 4, 8\}$.

We fix the samples size at 250 cases and 1000 controls for each simulated data set. The LSC model is applied to the high-quality measurements with 144 different prior configurations according to the search grid, then it returns a point estimate of the latent distribution under each prior configuration. The model is also applied to the low-quality measurements with the same grid of priors to study the impact of measurement quality. This procedure is repeated with independently simulated data sets for 25 times.

With 5 latent nodes, the returned latent distribution estimate is represented by a multinomial probability vector with 32 cells, denoted by $\widehat{q}_j$, $j = 1, \ldots, 32$. Let q_j be the true multinomial probabilities, then the Bhattachayya coefficient Bhattachayya (1943), $\sum_{j=1}^{32} \sqrt{q_j \widehat{q}_j} \in [0, 1]$, which measures the overlap of two discrete distributions, is used as the metric of the general estimation accuracy of the LSC model. To demonstrate the effectiveness of the LSC model, the partially-Latent Class (pLCM) model proposed by Wu, et al. (2015) Wu et al. (2015) is used as the benchmark. In this benchmark model, the latent nodes are parameterized by a multinomial distribution with 32 classes and a non-informative Dirichlet prior, and the prior choices for the TPRs and FPRs are the same as the LSC model. The posterior sample mean is then used as the point estimate of the latent distribution, whose accuracy is also evaluated using the Bhattacharyya Coefficient.

4.4.2 Results

Given each simulated data set and the pre-defined hyper-parameter search grid, 144 different prior configurations are evaluated. After each model fitting, we calculated the approximated DIC and the Bhattacharyya Coefficient (BC). The former is a measure of the goodness-of-fit, and the later is a measure of the latent distribution estimation accuracy. Figure 4.1 shows a few examples of how DIC and BC are correlated. As we can see, in each shown data set, based on either high or low-quality measurements, there is an apparent negative correlation between DIC and BC. In fact, the average percentile of the BC values that correspond to the lowest DIC of all tested priors is 80%. Thus, by choosing the prior configuration that yields the lowest DIC, we tend to pick from the models that produce the best estimation accuracy.

As a result of the prior selection procedure, table 4.3 lists the chosen prior configuration and the Bhattacharyya Coefficient for each simulated data sets with High-Quality measurements. Based on these 25 repetitions, the average BC resulted from the LSC model is 0.990 with standard deviation 0.004. In comparison, the benchmark model yields an average BC of 0.969 with standard deviation 0.012. Table 4.4 lists the chosen prior configuration and the Bhattacharyya Coefficient for each of the 25 simulated data sets with Low-Quality measurements. In this situation, the average BC value for the LSC model is 0.980 with standard deviation 0.010, but the average BC value for the benchmark model drops to 0.918 with standard

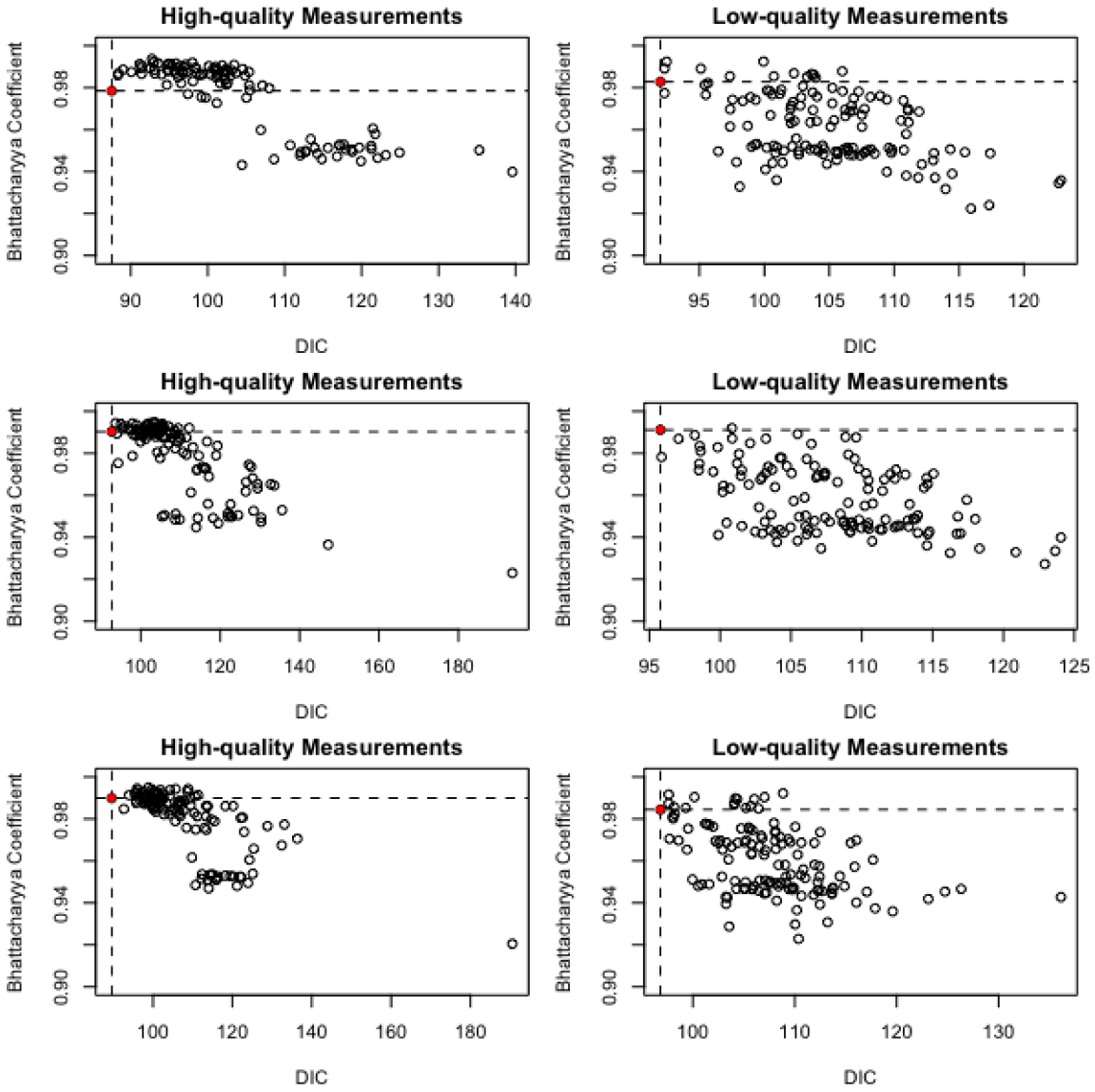

Fig. 4.1: **The Relation Between DIC and BC**: There are six graphs in this figure. Each row corresponds to a unique simulated data set (Here, three examples are randomly chosen from the 30 independent repetitions). The left column shows results based on high-quality measurements, and the right column is based on low-quality measurements. In each graph, the x-axis stands for the approximated DIC, and the y-axis stands for the Bhattacharyya Coefficient. Every dot in the graph corresponds to a unique prior configuration in the search grid. The model fitting result with the lowest DIC is labeled in red.

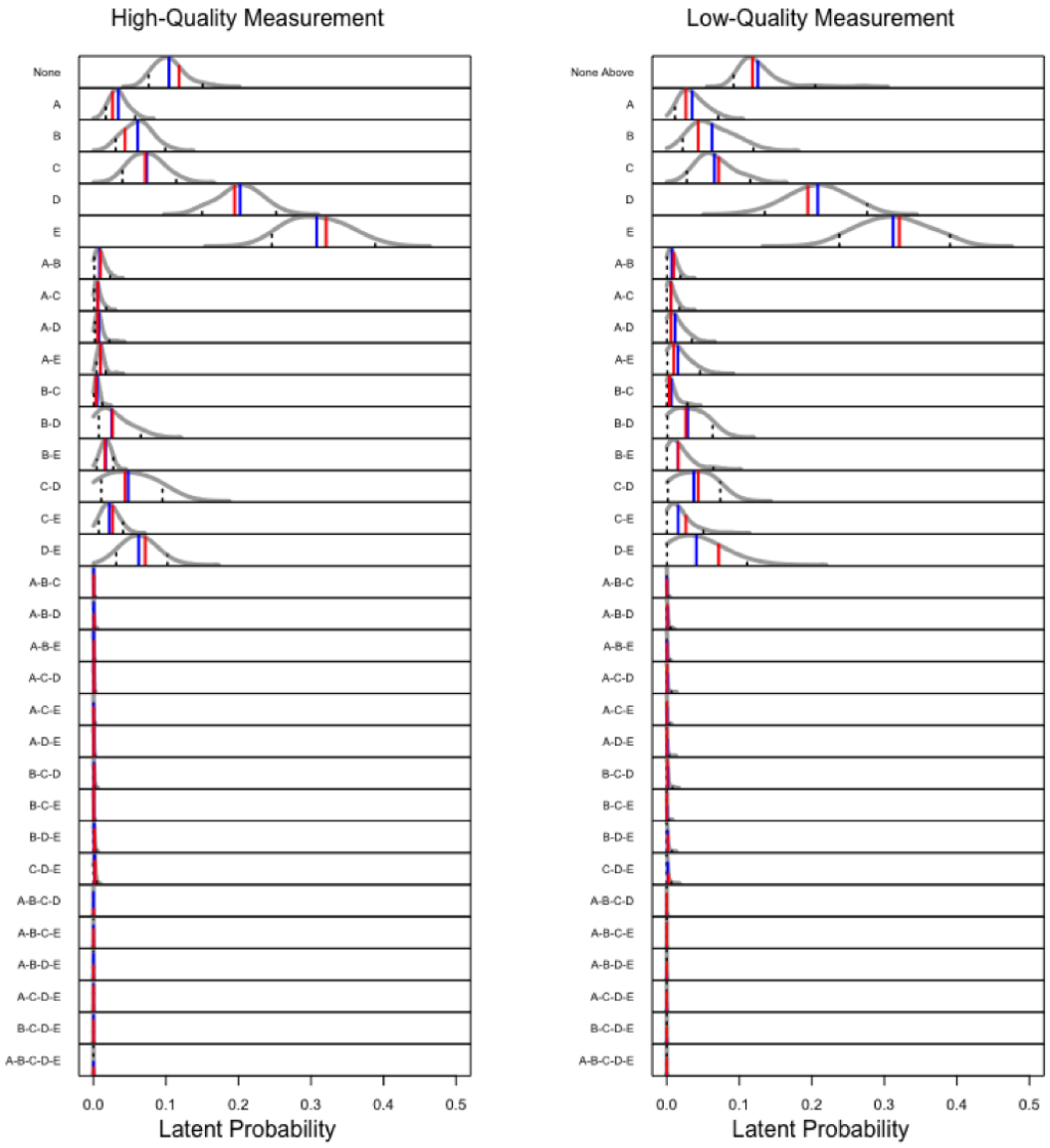

Fig. 4.2: **Latent Distribution Estimation**: The left plot is generated by model fitting results based on the High-Quality measurements, and the right plot is based on the Lower-Quality measurements. In each plot, the x-axis stands for probability value, and the y-axis includes the 32 unique combinations of the 5 latent nodes: A, B, C, D, E. These combinations are listed by the node names along the y-axis. For example, 'A-B' means only node $A = 1, B = 1$ and the rest nodes are 0. Each combination corresponds to a density curve of the sampling distribution of the probability of that latent node combination. Under each density curve, the red vertical line marks the true parameter value, and the blue vertical line shows the mean of the sampling distribution, then black dashed lines are the 2.5% and 97.5% percentiles of the sampling distribution.

deviation 0.050.

Table 4.3: Summary of High-Quality Measurements Analysis

Rep.ID	μ_ρ^*	τ_ρ^*	g_d	h_d	BC
1	-1	15	1	4	0.9907
2	-1	10	8	4	0.9892
3	-1	5	4	8	0.9784
4	-1	5	4	4	0.9903
5	-1	20	1	4	0.9912
6	-1	20	8	8	0.9873
7	-1	20	8	4	0.9938
8	-1	10	1	1	0.9928
9	-1	15	8	8	0.9900
10	-1	20	4	1	0.9890
11	-1	15	4	8	0.9906
12	-1	15	8	8	0.9948
13	-1	20	4	4	0.9899
14	-1	20	4	8	0.9810
15	-1	20	4	8	0.9908
16	-1	5	1	4	0.9917
17	-1	10	8	4	0.9904
18	-4	5	4	4	0.9912
19	-1	15	8	4	0.9866
20	-1	15	4	8	0.9945
21	-4	5	32	8	0.9908
22	-1	15	4	4	0.9914
23	-4	5	32	4	0.9931
24	-1	10	1	4	0.9854
25	-4	5	32	8	0.9921

Moreover, choosing from table 4.3 and 4.4, we fix the prior for high-quality data analysis at $(-1, 10, 1, 1)$, and fix the prior for low-quality data analysis at $(-4, 10, 32, 8)$, then we conduct another 150 independent repetitions for each analysis procedure. The resulting point estimates of the latent distribution serve as an approximation of the sampling distribution of our LSC estimator. These approxi-

Table 4.4: Summary of Low-Quality Measurements Analysis

Rep.ID	μ_ρ^*	τ_ρ^*	g_d	h_d	BC
1	-4	10	32	4	0.9756
2	-4	10	32	4	0.9828
3	-4	10	32	4	0.9911
4	-4	15	8	1	0.9659
5	-8	5	4	4	0.9682
6	-1	10	32	1	0.9733
7	-1	10	8	8	0.9733
8	-1	5	32	4	0.9668
9	-8	5	32	8	0.9844
10	-8	5	8	8	0.9800
11	-8	5	8	1	0.9876
12	-4	10	4	4	0.9897
13	-4	15	8	1	0.9902
14	-4	10	32	8	0.9545
15	-4	10	4	8	0.9861
16	-8	5	4	4	0.9927
17	-8	5	8	4	0.9871
18	-4	10	8	8	0.9772
19	-4	10	1	1	0.9891
20	-8	5	4	8	0.9804
21	-4	15	8	4	0.9901
22	-4	10	4	8	0.9880
23	-4	15	32	8	0.9883
24	-8	5	32	1	0.9890
25	-8	5	8	8	0.9777

mated sampling distributions are shown in figure 4.2. From the left figure, we can see that with high-quality measurements, the estimates returned by our proposed LSC model with the chosen prior configuration are mostly unbiased. Noticeable bias can be seen on Pr('None') and Pr('B') with absolute errors less than 0.018. In comparison, the right figure shows the results based on low-quality measurements, in which Pr('D-E') gets underestimated by 0.031 and the rest biases are close to the

Table 4.5: Summary of the Latent Distribution Estimates

Node Combination	True Value	HQ Estimates	HQ Std. Error	LQ Estimates	LQ Std. Error
None	0.118	0.104	0.02	0.125	0.029
A	0.026	0.034	0.011	0.035	0.016
B	0.043	0.061	0.018	0.062	0.027
C	0.072	0.074	0.021	0.066	0.023
D	0.195	0.202	0.028	0.208	0.038
E	0.321	0.308	0.039	0.312	0.044
A-B	0.01	0.008	0.006	0.007	0.005
A-C	0.006	0.006	0.004	0.006	0.005
A-D	0.006	0.008	0.005	0.011	0.01
A-E	0.01	0.01	0.004	0.015	0.013
B-C	0.004	0.005	0.003	0.006	0.007
B-D	0.026	0.025	0.018	0.029	0.021
B-E	0.016	0.017	0.006	0.015	0.015
C-D	0.043	0.048	0.029	0.037	0.024
C-E	0.026	0.022	0.009	0.016	0.015
D-E	0.072	0.062	0.019	0.041	0.031
A-B-C	0	0	0	0	0
A-B-D	0.001	0.001	0.001	0.001	0.001
A-B-E	0	0	0	0	0.001
A-C-D	0	0	0	0.001	0.002
A-C-E	0	0	0	0	0.001
A-D-E	0	0	0	0.001	0.001
B-C-D	0.001	0.001	0.001	0.001	0.002
B-C-E	0	0	0	0	0.001
B-D-E	0.001	0.001	0.001	0.001	0.002
C-D-E	0.002	0.002	0.001	0.001	0.002
A-B-C-D	0	0	0	0	0
A-B-C-E	0	0	0	0	0
A-B-D-E	0	0	0	0	0
A-C-D-E	0	0	0	0	0
B-C-D-E	0	0	0	0	0
A-B-C-D-E	0	0	0	0	0

high-quality analysis results. In addition to point estimates, the variational factors at convergence can be used to approximate the posterior variance as well. But it is known that such approximation tends to be an under-estimation. Specifically, the nominal 95% credible intervals only yield 9% to 43% actual coverage rates for the latent probabilities. Therefore, we recommend using resampling techniques to obtain interval estimates in practice.

The means and the standard deviations of the sampling distributions are listed in table 4.5. These results demonstrate that in both situations: high vs. low-quality measurements, the LSC model identifies the all the node combinations with non-zero probability. Also, it is expected to see that the standard errors based on low-quality measurements are mostly larger than those based on high-quality measurements. Under our simulation setting, the standard errors are increased by 58% on average.

4.5 Analysis of PERCH Data

PERCH study enrolled about 4200 children hospitalized with severe/very severe pneumonia and approximately 5300 controls randomly selected from communities across 7 sites around the world. To avoid having to adjust for the site-wise heterogeneity, only the Kenya site data, where there is good availability of both BS and SS measurement data, is used in the following analysis. We picked the top 10 etiologic pathogens, which could explain more than 90% of all infections according to Wu et al. (2015), as our candidate pathogens. These pathogens are streptococcus pneumoniae (PNEU), haemophilus influenzae (HINF), human metapneumovirus type A or B (HMPV_A_B), rhinovirus (RHINO), respiratory syncytial virus type A or B (RSV), parainfluenza type 1 virus (PARA_1), adenovirus (ADENO), Staphylococcus aureus (SAUR), coronavirus OC43 (COR), and influenza virus

type C (FLU_C). The BS measurements (nasopharyngeal specimen with PCR detection of pathogens - NPPCR) are available for all 281 cases and 1138 controls for all 10 pathogens. The SS measurements (blood culture results - BCX) are available for all cases, but only for the three bacteria pathogens: PNEU, HINF, and SAUR.

Prior scientific knowledge Murdoch et al. (2012) suggests that the TPR of BS measurements (NPPCR) is in the range of 50% - 99%, and the TPR of SS measurements (BCX) is in the range of 5% - 20%. Thus we set the hyper-parameters $a_k = 7.6$, $b_k = 59.0$, $c_k = 6.0$, and $d_k = 1.3$ by percentile matching, and $e_k = f_k = 1$ for non-informativeness. A binary feature, the disease severity ($0 =$ Severe, $1 =$ Very Severe), is used as the regression covariate, thus the etiology estimates can be obtained for two groups. Regression coefficients all have prior mean equal to zero and prior variance equal to 2. After a grid search for the hyper-parameters $(\mu_\rho^*, \tau_\rho^*, g_d, h_d)$, we identified that the configuration with the lowest DIC is $(\mu_\rho^* = -3, \tau_\rho^* = 10, g_d = 8, h_d = 2)$. Since Variational Bayesian methods tend to underestimate Wainwright et al. (2008) the variance of the posterior distribution, we approximate the standard error of our estimator by the basic nonparametric bootstrapping methodEfron (1992). That is, we randomly draw samples from the case and control data respectively with replacement to get a bootstrap sample of 281 cases and 1138 controls, and fit the LSC model to get a bootstrapped estimate, then the process is repeated for 200 times. The resulting estimates form a bootstrapped distribution for each parameter and its sample standard deviation is

considered the approximated standard error.

Figure 4.3 shows the bootstrapped distribution of the etiology probabilities for Kenya site, and table 4.6 lists the point estimates and approximated standard errors of them. As we can see, 10 to 15 percents of the infection could be caused by pathogens other than these ten. About 70 to 80 percents of all infections are singleton infections, where RSV contributes the most with roughly $20 - 30\%$. Another 10% are doubleton infections, and the rest about 990 different pathogen combinations only get to infect the lung with probability under 0.3%. Among all doubleton infections, only the RSV-HINF, PNEU-HINF, and PNEU-RHINO pairs contribute more than 1% in at least one severity group. These findings coincide with previous publications: Korppi et al. (1989) Korppi et al. (1989) find that RSV serves as a predisposing agent for secondary bacterial infection in the airways of children and HINF is one of the most common bacteria involved in the mixed RSV-bacterial infections in pneumonic patients. PNEU and HINF are both commensals of the human nasopharynx and have long been detected together in a multispecies biofilm in infected tissue Tikhomirova and Kidd (2013). Franz, Anna, et al. (2010) Franz et al. (2010) reported that in their study on lower respiratory tract infection, 28% RHINO infections has PNEU as the coinfected pathogen. According to the third column of figure 4.3, most of the bootstrapped distributions of the difference cover zero, while those of 'ADENO', 'SAUR', 'COR', 'RHINO-SAUR', and 'HINF-ADENO' deviate from zero to the left, suggesting that these etiologic combinations con-

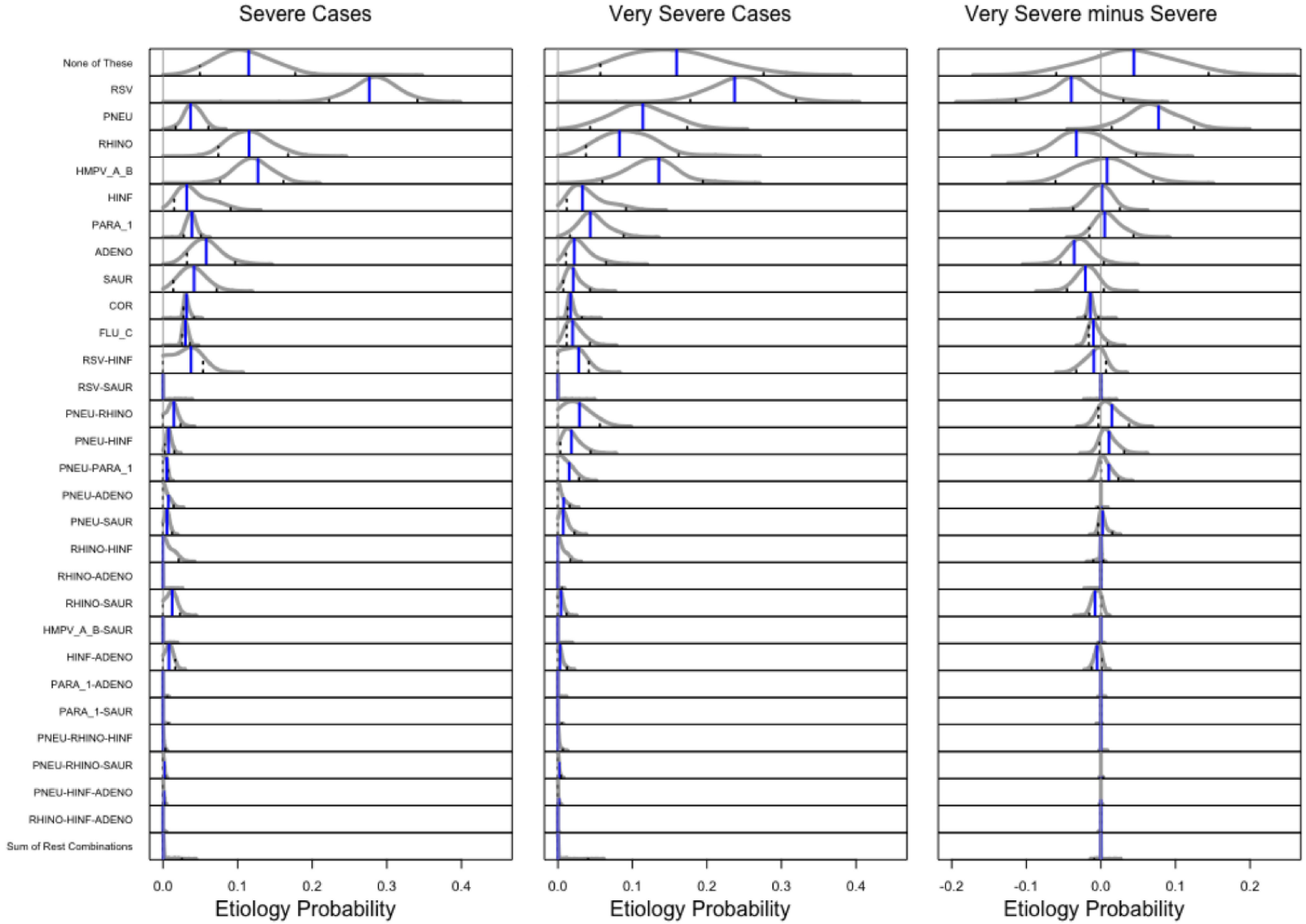

Fig. 4.3: **Bootstrapped Etiology Probability Estimation for Kenya site**: The left plot shows the results for the severe condition group, the middle plot shows the results for the very severe condition group, and the right plot visualizes the difference between the very severe and severe group. In each plot, the x-axis stands for the etiology probability value, and the y-axis includes the selected combinations of the 10 etiologic pathogens by their abbreviations. Only pathogen combinations whose 90th percentile of its bootstrapped etiology probability is greater than 10^{-4} are selected. Each combination corresponds to a density curve of the bootstrapped distribution of the etiology probability of that pathogen combination. Under each density curve, the blue vertical line represents the point estimate based on the original dataset, and the black dashed lines are the 2.5% and 97.5% percentiles of the bootstrapped

tribute more in the severe condition group. Also, those distributions of 'PNEU', 'PNEU-HINF', and 'PNEU-PARA_1' deviate to the right, which indicates that these infection patterns are more common in very severe cases.

Table 4.6: Etiology Probability Estimates for Kenya Site

Pathogen	Severe		Very Severe		Increment	
	Estimates	Std. Error	Estimates	Std.Error	Estimates	Std.Error
None of These	0.1096	0.0378	0.151	0.0586	0.0414	0.0548
RSV	0.2838	0.0289	0.2438	0.036	-0.0399	0.0343
PNEU	0.0401	0.0107	0.1102	0.0351	0.0701	0.0298
RHINO	0.1168	0.026	0.0943	0.0346	-0.0226	0.0329
HMPV_A_B	0.1203	0.0219	0.1252	0.0321	0.0049	0.0338
HINF	0.0417	0.0207	0.0391	0.0227	-0.0026	0.0165
PARA_1	0.0376	0.0057	0.0478	0.0183	0.0101	0.0161
ADENO	0.0572	0.0172	0.0295	0.0151	-0.0278	0.0152
SAUR	0.0394	0.0156	0.0207	0.0099	-0.0187	0.0133
COR	0.0316	0.0035	0.0181	0.0052	-0.0135	0.0041
FLU_C	0.0299	0.003	0.0219	0.0099	-0.0081	0.0072
RSV-HINF	0.0255	0.0196	0.0171	0.0143	-0.0084	0.0109
RSV-SAUR	0.00417	0.0093	0.00239	0.00641	-0.00179	0.00525
PNEU-RHINO	0.01064	0.00701	0.02244	0.01708	0.0118	0.01227
PNEU-HINF	0.0078	0.0033	0.0175	0.0106	0.0097	0.0092
PNEU-PARA_1	0.0024	0.0024	0.0082	0.009	0.0058	0.007
PNEU-ADENO	0.003	0.0045	0.0031	0.0049	1e-04	0.0019
PNEU-SAUR	0.0049	0.0033	0.0074	0.006	0.0025	0.0046
RHINO-HINF	0.005	0.0073	0.0037	0.0056	-0.0014	0.0033
RHINO-ADENO	0.0022	0.0055	7e-04	0.0018	-0.0015	0.0039
RHINO-SAUR	0.0096	0.007	0.0038	0.0034	-0.0058	0.0051
HMPV_A_B-SAUR	0.0018	0.0044	0.0018	0.0044	0	9e-04
HINF-ADENO	0.0074	0.005	0.0033	0.0034	-0.004	0.0038
PARA_1-ADENO	6e-04	0.0017	8e-04	0.0023	2e-04	0.001
PARA_1-SAUR	0.0012	0.0023	6e-04	0.0014	-5e-04	0.0013
PNEU-RHINO-HINF	5e-04	9e-04	0.001	0.002	5e-04	0.0014
PNEU-RHINO-SAUR	7e-04	9e-04	8e-04	0.0012	1e-04	8e-04
PNEU-HINF-ADENO	5e-04	8e-04	4e-04	7e-04	-1e-04	4e-04
RHINO-HINF-ADENO	2e-04	7e-04	1e-04	2e-04	-2e-04	5e-04
Sum of Rest Combinations	0.0035	0.0076	0.0032	0.0094	-4e-04	0.0046

4.6 Discussion

In this paper, we propose a fast posterior approximation method for the Bayesian latent sparse correlation model (BLSCM), with primary application to pneumonia etiology estimation using non-Gold standard measurements. By using the mean-field variational family, we develop a coordinate acsent variational inference (CAVI) algorithm, which is extraordinarily fast and scalable. Let K be the dimension of the latent state and let n be the sample size. In each iteration, the algorithm only requires $O(K^2n)$ arithmetic calculations and $O(K)$ unconstrained univariate optimizations. Also, a DIC-based procedure is also proposed for selecting optimal hyper-parameters. It is shown by simulation studies that with the proposed prior selection procedure, we are able to pick among models with the best estimation accuracy. Although the problem is only partially identifiable and the estimation is expected to be biased towards the prior, the estimation produced by the chosen model for the latent distribution is reasonably accurate, reaching a higher Bhattacharyya Coefficient than the benchmark model does. Also, as the measurement quality gets worse, the estimation accuracy of the LSC model declines much slower than the benchmark model does. In terms of actual computing time, a 10-dimensional (10 latent binary nodes) problem with 300 cases used to take a few minutes to finish a single iteration with an MCMC algorithm or even longer with a conventional EM algorithm. With the CAVI algorithm, the whole estimation proce-

dure from start to convergence only takes about 10 seconds on a 2.5GHz CPU. In the PERCH data analysis, the estimation results are consistent with published etiology research findings and provide insights into the interactions of childhood pneumonia etiologic pathogens and how the infection patterns differ between two severity levels.

A shortcoming of the mean-field variational approximation is that the bias of its posterior predictive value could be large when the posterior dependency between variables are strong, and the posterior variances are often underestimated. Our proposed estimating procedure has to rely on resampling method to approximate the standard error. Thus, future work could involve exploring different variational families that allow for interactions between factors, such as the partially factorized structure (Saul and Jordan, 1996) and the decimatable Boltzmann machine (Barber and Wiegerinck, 1999), so that the both the posterior mean and variance can be better approximated. Also, since there is not much theory developed for the general asymptotic behavior of variational inference, an important extension of this work is to show the theoretical guarantees of the variational approximation of the Bayesian latent sparse correlation model.

Chapter 5

Efficient Estimation of Time-to-event Models by Incorporating Auxiliary Survival Information

CHAPTER 5. EFFICIENT ESTIMATION OF TIME-TO-EVENT MODELS BY INCORPORATING AUXILIARY SURVIVAL INFORMATION

Abstract

For the purpose of customizing treatment for each individual patient, the primary task is to evaluate the treatment effects in different sub-populations, i.e. sub-group analysis. With the knowledge of the expected treatment effect in each sub-population, clinicians can simply match individual patient to a sub-group by its profile and adopt the best intervention. A major challenge in sub-group analysis is the lack of statistical power. Combining auxiliary information into individual clinical analysis is an emerging strategy to improve estimation efficiency. We propose a novel estimating procedure for improving the efficiency of survival models by incorporating external information on the population level survival rates. The accelerated failure time (AFT) model and the Cox proportional hazards model are considered. For each model, we describe how we derive the set of over-identifying moment conditions from the benchmark estimators and auxiliary information. Then, the parameter estimation and model diagnostics are carried out following the standard generalized method of moments (GMM) framework. We show that the our GMM-based estimators are asymptotically and empirically more efficient than the benchmark estimators. These new estimators are applied to a recent retrospective study on the prognosis factors of pancreatic cancer.

5.1 Introduction

Individualized health aims to provide disease treatment and prevention based on individual characteristics of the genome, medical imaging, family history, environment, and lifestyle. The core of customizing treatment for each patient is to evaluate the treatment effects in different sub-populations, i.e., sub-group analysis. With the knowledge of the expected treatment effect in each sub-population, clinicians can simply match the individual patient to a sub-group by its profile and adopt the best intervention. A significant challenge in sub-group analysis is the lack of statistical power. For example, there are two types of androgen deprivation therapy (ADT) for men with advanced prostate cancer. Continuous ADT (CADT) is the conventional treatment in the US, and intermittent ADT (IADT) is proposed as an alternative treatment with potential benefits regarding the quality of life, financial cost, and side effects. Moreover, as age and prostate-specific antigen (PSA) level are reportedly the key prognostic factors for advanced prostate cancer, it is of great interest to evaluate the potential advantage of IADT over CADT, and especially to examine whether such potential effect differs by PSA level or age at the time of diagnosis. However, the recent clinical study (Hussain et al., 2013) was not able to prove/disprove the comparative effectiveness of IADT versus CADT in each group of PSA level and age. Therefore, techniques for improving the statistical efficiency are of critical value for sub-group analysis and individualized treatment.

In recent years, the availability of extremely large datasets is increasing rapidly. Such datasets include but not limit to population census data, disease registries and electronic health records. For example, the Surveillance, Epidemiology, and End Results (SEER) Program (19732014) provides cancer incidence and survival data from cancer registries that approximately covers 30% of the US population. These datasets provide a valuable source of information that can be utilized to improve the design and analysis of individual studies (Gail et al., 1989; Costantino et al., 1999; Wu and Sitter, 2001; Chatterjee et al., 2015). Inspired by these works, we are interested in developing estimating procedures to improve the statistical efficiency of individual studies by incorporating auxiliary information from large external datasets. In the rest of this paper, we will refer to such estimating procedure as evidence synthesis, or information synthesis. Admittedly, the form of information available in these larger datasets varies a lot. At one extreme, some datasets grant public access to individual-level data, while at the other extreme, only population-level summaries are available. We focus on the latter situation, that is, utilizing the population-level summary statistics, since patient-level data are typically not available to the public due to ethical concerns.

To this end, we have considered a few approaches for information synthesis. Imbens and Lancaster (1994) showed that in economic studies, cross-sectional or panel samples can be combined with the population moments of the economic variables extracted from census reports using the generalized method of moments

(GMM) to improve estimation accuracy. Qin (2000) showed that empirical likelihood (EL) can be used for combining auxiliary information with any parametric likelihood as long as such information can be expressed as a set of unbiased estimating equations, and Qin et al. (2014) developed an empirical likelihood (Owen, 2001) approach to utilize the covariate-specific disease prevalence information to improve the efficiency of logistic regression with case-control data. Recently, Chatterjee et al. (2015) summarized previous works and proposed a general semi-parametric maximum likelihood estimation methodology. In this approach, the auxiliary information is provided as a finite set of parameters resulted from fitting a model to the external large dataset, regardless of whether the model is correctly specified. The external model is used to identify a set of constraints that link the individual-level data to the auxiliary information, which leads to the constrained maximum likelihood estimation.

Essentially, all these approaches share the same idea, which is to impose additional constraints identified from auxiliary information on the original model. These constraints are typically converted into unbiased estimating equations (moment conditions). Then, different techniques based on the generalized method of moments (GMM) or the empirical likelihood (EL) method are proposed for parameter estimation under additional constraints. Since first formalized by Hansen (1982), GMM has been an important and frequently used estimation technique for many econometric and quantitative finance problems, such as estimating the structural

model in macroeconomics and the capital pricing model in finance. The asymptotic as well as finite-sample properties of GMM are also extensively studied (Newey and West, 1987; Pakes and Pollard, 1989). Under mild regularity conditions, GMM estimator is shown to be consistent (Hansen, 1982) and asymptotically normal. More importantly, the efficiency typically gets improved as the number of moment conditions goes up. Suppose there are p parameters and q moment conditions, the problem is termed under-identified if $q < p$, just-identified if $q = p$, and over-identified if $q > p$. Essentially, the GMM approach (Imbens and Lancaster, 1994) improves the estimation efficiency by constructing over-identifying moments derived from the likelihood as well as the auxiliary information. The empirical likelihood method (Owen, 2001) is another flexible estimating procedure based on moment conditions. Assuming the set of over-identifying moments are correctly formulated, Qin (2000) showed that efficiency gain could also be reached with the empirical likelihood approach. In fact,Qin and Lawless (1994) and Imbens (2012) proved that the EL estimator is asymptotically equivalent to the two-step GMM (Hansen, 1982) estimator. Moreover, Newey and Smith (2004) and Anatolyev (2005) showed that the EL estimator has smaller second order bias than GMM does, and the bias does not grow as the number of over-identifying restrictions increases. But Guggenberger (2008) argued that such asymptotic advantages of EL over GMM may not hold in small sample problems.

Recent works on information synthesis in survival analysis with right-censoring

provide several examples of adopting or extending the above methods to different types of regression models. Zhou (2006) adopted the empirical likelihood framework to improve the efficiency of Cox model with partially known baseline hazard. Huang et al. (2015) developed an empirical likelihood estimator of Cox model that incorporates auxiliary subgroup survival probabilities. Under the additive−multiplicative hazard assumption, GMM based estimator that utilizes auxiliary survival information is developed in Shang and Wang (2017). However, for the accelerated failure time (AFT) model, information synthesis is still an open question. The two most important semi-parametric estimating procedures for AFT model are the rank-based approach (Jin et al., 2003) and the least-square approach (Jin et al., 2006). Both approaches make no assumption on the error distribution, and both estimations are consistent and asymptotically normal. Neither approach is uniformly more efficient than the other. Simulation studies show that the log-rank estimator is more efficient if the errors are simulated from an extrem-value distribution, while the least-square estimator is more efficient with normal errors (Jin et al., 2006). We focus on the log-rank estimator with Gehan-type weights for its computational advantage.

To improve the efficiency of this log-rank estimator through information synthesis, we propose a novel way of adopting the GMM framework to incorporate the auxiliary sub-group survival information in the AFT model. We approach this problem from the perspective of GMM rather than EL primarily because of GMM's computational advantage: (1) The standard two-step GMM estimator requires two

optimization steps over a p-dimensional space, while the EL estimator requires maximization over a $p + n$-dimensional space, where n is the sample size, subject to $q + 1$ restrictions. The later one is in general a more formidable task. (2) Huang et al. (2015) reported that the EL optimizer tends to be computationally unstable when multiple landmark time survival probability constraints are used, since these constraints tend to be highly correlated. Adopting Imbens and Lancaster (1994)'s framework to right-censored survival data is a challenging task. Under the GMM framework, the moment conditions are typically considered as an average of i.i.d terms, $\frac{1}{n}g(x_i;\theta)$, where $x_1, x_2, \ldots, x_n$ are independent samples and θ is the parameter of interest. But with right-censored data, each summand g_i in the unbiased estimating function is dependent on the full dataset, which creates bias in the estimation of the covariance matrix of the moment conditions. Shang and Wang (2017) proposed a modified two-step estimator for this problem and showed that the solution is consistent and asymptotically normal. But the well-studied properties of the GMM estimators (two-step, iterative, continuously updated) no longer apply. For example, the Sargan-Hansen J-statistic will not converge to a Chi-square distribution, thus the J-test for checking the compatibility of the over-identifying conditions becomes invalid. In this work, we propose to use the functional delta method (Van der Vaart, 1998) to transform the estimating functions to the typical moment condition form, an average of asymptotically i.i.d terms. Then these transformed moments can be fed into the usual GMM estimating procedure directly, and the

theories of GMM estimator would all apply.

5.2 Methods

In this section, we will introduce how we derive the over-identifying moment conditions from the log-rank estimator of the AFT model as well as the auxiliary sub-population survival information. Then we will show that the our GMM estimator is asymptotically more efficient than the log-rank estimator without information synthesis. In addition, we will provide an extended of the GMM estimator so that it can adjusts for the possibly inconsistent baseline hazards between the auxiliary information and the individual dataset. Apart from the primary contribution above, we will also demonstrate how we can apply the same approach to develop an efficient GMM estimator for the Cox proportional hazards model.

5.2.1 Notations and Terminology

Let T be the failure time, that is, time to the event of interest. Denote C as the censoring time, and X as a vector of baseline covariates. The observed time is denoted by $Y = \min(T, C)$ and the censoring indicator is $\Delta = I(T \leq C)$. Assume that C is conditionally independent of T given X. Essentially, our approach can be applied to combine any form of auxiliary information as long as it can be represented by a set of moment conditions. We focus on utilizing the sub-population

survival rates because it is highly relevant to survival analysis, and it is often publicly available in disease registries. For example, in the Surveillance, Epidemiology, and End Results (SEER) Program, the SEER Cancer Statistics Review 1973-2014 (Horner et al., 2015) reports the 5-year survival rates by race, sex, age, and year of diagnosis for the major cancer sites and for all cancers combined using data from 18 cancer registries, which cover around 28$\%$ of the US population. Suppose the disease registries keep records of the t^*-year survival rates for K different sub-groups of the diseased population. The group membership is determined by the covariate information in the disease registries. Let $\Omega_k, k = 1, \ldots, K$ be the kth sub-group, then the auxiliary information at a landmark time point t^* is $P(T > t^*|X \in \Omega_k) = \phi_k$. For example, let Z_1 denote the age at diagnosis, and Z_2 be the biomarkers and risk factors of ovarian cancer that are not available in the SEER database. With the complete covariate $X = (Z_1, Z_2)$, the auxiliary survival information for ovarian cancer patients according to the SEER program can be represented as $P(T > 5|X \in \Omega_1) = 0.56$ and $P(T > 5|X \in \Omega_2) = 0.277$, where $\Omega_1 = \{(Z_1, Z_2) : Z_1 < 65\}$, and $\Omega_2 = \{(Z_1, Z_2) : Z_2 \geq 65\}$.

5.2.2 GMM Estimator for the AFT Model

5.2.2.1 Moment Conditions

For $i = 1, 2, \ldots, n$, we assume that, given the covariates $X_i = x$, the survival time T_i follows the AFT model $\log(T_i) = x^T\beta + \epsilon_i$, where β is a vector of the regression coefficients and the log baseline survival time ϵ_i has an unspecified distribution. We define the log-scale residual $e_i(\beta) = \log Y_i - X_i^T\beta$, and $N_i(\beta, t) = \Delta_i I(e_i(\beta) \leq t)$. Note that N_i is the counting process on the time scale of the residual. The log-rank estimating function for β with the Gehan-type weight is given by:

$$\begin{aligned}\Phi(\beta) =& \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} S^{(0)}(\beta, u)[X_i - \frac{S^{(1)}(\beta, u)}{S^{(0)}(\beta, u)}]dN_i(\beta, u) \\ =& \frac{1}{n}\sum_{i=1}^{n}\Big[\Delta_i S^{(0)}(\beta, e_i)X_i - \Delta_i S^{(1)}(\beta, e_i)\Big], \end{aligned} \tag{5.1}$$

where $S^{(0)}(\beta, t) = \frac{1}{n}\sum_{j=1}^{n} I(e_j(\beta) \geq t)$, $S^{(1)}(\beta, t) = \frac{1}{n}\sum_{j=1}^{n} I(e_j(\beta) \geq t)X_j$. Define $s_k(\beta, t) = \mathbb{E}[S^{(k)}(\beta, t)]$, $k = 0, 1$, where $\mathbb{E}(\cdot)$ is the expectation over the true joint distribution of (T, C, X). Let $\widehat{\beta}_G$ be a root of the estimating function $\Phi(\beta) = 0$, and β_0 be the true value of the regression coefficient. Under regularity conditions: (a) The covariates are uniformly bounded. (b) The density function of the error distribution f and its gradient f' are bounded and $\int (f'(t)/f(t))^2 f(t)dt < \infty$. (c) The log censoring time $\widetilde{C}_i$ have uniformly bounded densities for all t and i. (d)

$\sup_i \mathbb{E}|\min\{\epsilon_i, \widetilde{C}_i\}|_0^\gamma < \infty$, for some $\gamma_0 > 0$. (e) The space of β is a compact set and β_0 is in its interior. **?** showed that almost surely

$$\Phi(\widehat{\beta}_G) = \Phi(\beta_0) + D_G(\widehat{\beta}_G - \beta_0) + o(1/\sqrt{n} + ||\widehat{\beta}_G - \beta_0||), \tag{5.2}$$

and

$$\sqrt{n}(\widehat{\beta}_G - \beta_0) \xrightarrow{D} N(0, D_G^{-1}\Sigma D_G^{-1}), \tag{5.3}$$

where

$$D_G = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} S^{(0)}(\beta, u)[X_i - \frac{S^{(1)}(\beta_0, u)}{S^{(0)}(\beta_0, u)}]^{\otimes 2} \frac{d\lambda(u)/du}{\lambda(u)} dN_i(\beta_0, u), \tag{5.4}$$

$$\Sigma = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} S^{(0)}(\beta, u)^2[X_i - \frac{S^{(1)}(\beta_0, u)}{S^{(0)}(\beta_0, u)}]^{\otimes 2} dN_i(\beta_0, u) = \mathsf{var}[\sqrt{n}\Phi(\beta_0)], \tag{5.5}$$

and $\lambda(\cdot)$ is the hazard function of ϵ_i.

To transform estimating function 5.1 to a summation of i.i.d terms, we apply the

functional delta method (Van der Vaart, 1998) to $\Phi(\beta)$, then

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\int S^{(0)}(\beta,u)X_i dN_i(\beta,u) \qquad (5.6)\\ =&\frac{1}{n}\sum_{i=1}^{n}\Big\{\int[I(e_i(\beta)\geq u)-s_0(\beta,u)]\mathbb{E}[XdN(\beta,u)]+\int s_0(\beta,u)X_i dN_i(\beta,u)\Big\}+o_p(\frac{1}{\sqrt{n}}),\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\int S^{(1)}(\beta,u)dN_i(\beta,u) \qquad (5.7)\\ =&\frac{1}{n}\sum_{i=1}^{n}\Big\{\int[I(e_i(\beta)\geq u)X_i-s_1(\beta,u)]\mathbb{E}[dN(\beta,u)]+\int s_1(u)dN_i(\beta,u)\Big\}+o_p(\frac{1}{\sqrt{n}}).\end{aligned}$$

Plugging formula 5.6 and 5.7 into 5.1, we have the asymptotically i.i.d representation $\Phi(\beta)=\frac{1}{n}\sum_{i=1}^{n}g_i^{(1)}(\beta)+o_p(\frac{1}{\sqrt{n}})$, where

$$\begin{aligned}g_i^{(1)}(\beta)=&\int[I(e_i(\beta)\geq u)-s_0(\beta,u)]\mathbb{E}[XdN(\beta,u)]+\int[s_0(\beta,u)X_i-s_1(\beta,u)]dN_i(\beta,u)\\ &-\int[I(e_i(\beta)\geq u)X_i-s_1(\beta,u)]\mathbb{E}[dN(\beta,u)], \qquad (5.8)\end{aligned}$$

and $g_i^{(1)}(\beta)$ is independent with $g_{i'}^{(1)}(\beta)$ for $i\neq i'$. Thus, $\Sigma=\text{var}\Big[g_i^{(1)}(\beta_0)\Big]=\mathbb{E}[g_i^{(1)}(\beta_0)g_i^{(1)}(\beta_0)^T]+o_p(1)$, which can be consistently estimated by

$$\widehat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}\Big[g_i^{(1)}(\widehat{\beta}_G)-\frac{1}{n}\sum_{j=1}^{n}g_j^{(1)}(\widehat{\beta}_G)\Big]\Big[g_i^{(1)}(\widehat{\beta}_G)-\frac{1}{n}\sum_{j=1}^{n}g_j^{(1)}(\widehat{\beta}_G)\Big]^T.$$

Denote $\sum_{i=1}^{n} g_i^{(1)}(\beta)$ as $\Phi^*(\beta)$, then $\widehat{\beta}_G$ asymptotically solves $\Phi^*(\beta) = 0$ and

$$\Phi^*(\widehat{\beta}_G) = \Phi^*(\beta_0) + D_G(\widehat{\beta}_G - \beta_0) + o(1/\sqrt{n} + ||\widehat{\beta}_G - \beta_0||) + o_p(1/\sqrt{n}), \tag{5.9}$$

that is, estimating functions $\Phi(\beta)$ and $\Phi^*(\beta)$ have the same asymptotic slope matrix D_G. This i.i.d representation $\Phi^*(\beta)$ is used as the first set of moment restrictions for our GMM estimator.

Next, we construct the second set of moments that incorporates the auxiliary survival information. First, define α_i as the conditional cumulative hazard at time t^* given X_i. Applying the Nelson-Aalen estimator (Park and Wei, 2003), we have

$$\widehat{\alpha}_i(\beta, t^*) = \frac{1}{n} \sum_{j=1}^{n} \int_{-\infty}^{\log t^* - X_i^T \beta} \frac{1}{S^{(0)}(\beta, u)} dN_j(\beta, u).$$

Then by applying the functional delta method, we can show that

$$\begin{aligned} \widehat{\alpha}_i(\beta, t^*) = & \frac{1}{n} \sum_{j=1}^{n} \Big\{ \int_{-\infty}^{\log t^* - X_i^T \beta} \Big[\frac{1}{s_0(\beta, u)} - \frac{I(e_j(\beta) \geq u)}{s_0(\beta, u)^2} \Big] \mathbb{E}[dN(\beta, u)] + \\ & \int_{-\infty}^{\log t^* - X_i^T \beta} \frac{1}{s_0(\beta, u)} dN_j(\beta, u) \Big\} + o_p(\frac{1}{\sqrt{n}}). \end{aligned} \tag{5.10}$$

Therefore, the individual survival probability estimate at time t^* is $\exp\{-\widehat{\alpha}_i(\beta, t^*)\}$. Recall that the auxiliary information is given as the sub-population survival probability at a landmark time point t^*: $P(T > t^* | X \in \Omega_k) = \phi_k$. By double expectation, the individual survival information at time t^* amounts to the population moments

$\mathbb{E}\{\Psi_k(X_i,\beta)\}=0$, where for $k=1,\ldots,K$:

$$\Psi_k(X_i,\beta)=I(X_i\in\Omega_k)[\exp\{-\widehat{\alpha}_i(\beta,t^*)\}-\phi_k].$$

Plugging in $\widehat{\alpha}_i$, the moment condition derived from the auxiliary information becomes

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Psi_k(X_i,\beta)=&\frac{1}{n}\sum_{i=1}^{n}I(X_i\in\Omega_k)[\exp\{-\widehat{\alpha}_i(\beta,t^*)\}-\phi_k]\\ =&\frac{1}{n}\sum_{i=1}^{n}I(X_i\in\Omega_k)\Big\{(1+\alpha_i)e^{-\alpha_i}-\phi_k+o_p(\frac{1}{\sqrt{n}})\Big\}-\\ &\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}I(X_i\in\Omega_k)e^{-\alpha_i}Q_{ij}(\beta),\end{aligned}\tag{5.11}$$

$$\begin{aligned}\text{where}\quad Q_{ij}(\beta)=&\int_{-\infty}^{\log t^*-X_i^T\beta}\Big[\frac{1}{s_0(\beta,u)}-\frac{I(e_j(\beta)\geq u)}{s_0(\beta,u)^2}\Big]\mathbb{E}[dN(\beta,u)]+\\ &\int_{-\infty}^{\log t^*-X_i^T\beta}\frac{1}{s_0(\beta,u)}dN_j(\beta,u).\end{aligned}\tag{5.12}$$

Note that this is still not in the independent summation form we seek, rather we get a nested summation form, which can be regarded as a U-statistic. This motivates us to apply the following approximation. Let $d_i=(y_i,x_i,\Delta_i)$ be the observed data vector for subject i. Define the kernel function $q(d_i,d_j)=q_{ij}=I(X_i\in\Omega_k)e^{-\alpha_i}Q_{ij}(\beta)+I(X_j\in\Omega_k)e^{-\alpha_j}Q_{ji}(\beta)$, then let $R=\frac{1}{\binom{n}{2}}\sum_{i<j}q_{ij}$. Applying the Hajek

projection of U-statistics (Hájek, 1968), we can show that

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} I(X_i \in \Omega_k)e^{-\alpha_i}Q_{ij}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_j}q(d_i, D_j) - \frac{1}{2}\mathbb{E}(R) + \frac{1}{2n^2}\sum_{i=1}^{n} q_{ii} + o(\frac{1}{n^2}) + o_p(\frac{1}{\sqrt{n}}). \tag{5.13}$$

Plugging back to Ψ_k yields

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_k(X_i, \beta) = \frac{1}{n}\sum_{i=1}^{n} g_{ik}^{(2)}(\beta) + o(\frac{1}{n^2}) + o_p(\frac{1}{\sqrt{n}}),$$

where

$$g_{ik}^{(2)}(\beta) = I(X_i \in \Omega_k)\big[(1+\alpha_i)e^{-\alpha_i} - \phi_k - \frac{1}{n}e^{-\alpha_i}Q_{ii}(\beta)\big] - \mathbb{E}_{D_j}q(d_i, D_j) + \frac{1}{2}\mathbb{E}(R), \quad \text{for } k = 1, \ldots, K. \tag{5.14}$$

Each $g_{ik}^{(2)}(\beta)$ has a zero mean with finite second moment, and $g_{ik}^{(2)}(\beta)$ is independent with $g_{i'k}^{(2)}(\beta)$ for $i \neq i'$. As a result, the log-rank estimating function as well as the auxiliary information are transformed into a set of over-identifying moment conditions, $\frac{1}{n}\sum_{i=1}^{n} G_i(\beta)$, where $G_i(\beta) = (g_i^{(1)T}(\beta), g_{i1}^{(2)T}(\beta), \ldots, g_{iK}^{(2)T}(\beta))^T$.

5.2.2.2 Estimation and Inference

After formulating the moment conditions as $\frac{1}{n}\sum_{i=1}^{n} G_i(\beta)$, we can apply the standard GMM (Hansen, 1982; Imbens and Lancaster, 1994) estimating procedure to obtain a consistent and asymptotically efficient estimators. Specifically, we estimate the true regression coefficient β_0 by minimizing the following loss function:

$$L(\beta) = \Big[\frac{1}{n}\sum_{i=1} G_i(\beta)\Big]^T W \Big[\frac{1}{n}\sum_{i=1} G_i(\beta)\Big], \tag{5.15}$$

where W is the weight matrix, and the optimal weight is given by var$[\frac{1}{n}\sum_{i=1} G_i(\beta_0)]^{-1}$, that is, the inverse of the covariance matrix of the moment conditions evaluated at the true value. The loss function 5.15 is minimized using the iteratively update algorithm described in Hansen et al. (1996), that is: (1) Minimize $L(\beta)$ with W set as an identity matrix and denote the minimizer as $\beta^{(0)}$. (2) Update $W =$ var$[\frac{1}{n}\sum_{i=1} G_i(\beta^{(0)})]^{-1}$. (3) Minimize $L(\beta)$ with updated W and obtain $\beta^{(1)}$. (4) If $||\beta^{(1)} - \beta^{(0)}||$ is smaller than tolerance, then define $\widehat{\beta} = \beta^{(1)}$ as our GMM estimate with information synthesis. Otherwise, set $\beta^{(0)} = \beta^{(1)}$ and go back to step (2). The asymptotic property of the AFT-GMM estimator $\widehat{\beta}$ is summarized in the theorem 2.1.

Theorem 5.2.1. *Under regularity conditions (a) to (e), and conditions in Newey and*

McFadden (1994) Theorem 7.2, we have:

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{D} N(0, \Gamma^{-1}),$$

where $\Gamma^{-1} = (D_G \Sigma^{-1} D_G + BA^{-1}B)^{-1}$*, provided* D_G *is non-singular, where* A *and* B *are defined in the Appendix.*

As we can see, theorem 2.1 implies that the AFT-GMM estimator that combines the information from both the subject-level data and the auxiliary sub-population survival information is asymptotically more efficient than the original log-rank estimator with Gehan's weight. In practice, since the moment functions are possibly non-smooth, we apply the random perturbation algorithm proposed by Chen and Liao (2015) to estimate the asymptotic variance-covariance matrix.

The validity of the AFT-GMM estimator requires the auxiliary survival information to be consistent with the individual-level data. To check whether this consistency condition holds, we can apply the Sargan-Hansen J-test. Let q be the number of moment conditions (the dimension of vector $G(\beta)$), and p be the dimension of parameters. Under the regularity conditions specified in theorem 2.1, and under the null hypothesis (H_0) that $\mathbb{E}(\frac{1}{n}\sum_{i=1} G_i(\beta_0)) = 0$, we have

$$J := n\Big[\frac{1}{n}\sum_{i=1} G_i(\widehat{\beta})\Big]^T \widehat{W_n}\Big[\frac{1}{n}\sum_{i=1} G_i(\widehat{\beta})\Big] \xrightarrow{D} \chi^2_{q-p},$$

where $\widehat{W_n}$ is the iteratively updated weight matrix that converges to the optimal

weight matrix. With the alternative hypothesis (H_1) being $\mathbb{E}(\frac{1}{n}\sum_{i=1} G_i(\beta)) \neq 0$ for any β value in the domain, we can reject the null that the auxiliary information is consistent with the individual-level data with 95% confidence, if $J > q_{0.95}^{\chi^2_{q-p}}$.

5.2.2.3 Account for the Inconsistency in Baseline Hazard

Sometimes, due to the differences in the study inclusion/exclusion criteria, subjects enrolled in an individual clinical study may not be a representative sample of the population where we extract the auxiliary survival information from. In such cases, the information in the individual-level data could be inconsistent with the auxiliary information. Thus, we propose an extended AFT-GMM estimator to account for the potential inconsistency in the baseline hazard function from the two sources. Essentially, the extended estimator allows the auxiliary baseline hazard function $\lambda_0^a(t)$ to be different than the baseline hazard function of the study population $\lambda_0(t)$ by a scaling factor ρ, that is, for any t

$$\lambda_0^a(t) = \rho\lambda_0(t), \rho > 0. \tag{5.16}$$

Intuitively, $\rho = 1$ suggests that the information in the individual-level data is inconsistent with the auxiliary information, while $\rho > 1$ indicates that the study population has a smaller baseline hazard rate comparing to the population where the auxiliary information is drawn. As a result, the auxiliary survival information can be

summarized by $\frac{1}{n}\sum_{i=1}^{n}\Psi_k(X_i,\beta,\rho)$, where

$$\Psi_k(X_i,\beta,\rho) = I(X_i \in \Omega_k)[\exp\{-\rho\widehat{\alpha}_i(\beta,t^*)\} - \phi_k].$$

Applying the same argument we used in section 2.2.1 yields the following moment condition for the auxiliary information,

$$\begin{aligned} g_{ik}^{(2)}(\beta,\rho) =& I(X_i \in \Omega_k)\big[(1+\rho\alpha_i)e^{-\rho\alpha_i} - \phi_k - \frac{1}{n}\rho e^{-\rho\alpha_i}Q_{ii}(\beta)\big] \\ & -\mathbb{E}_{D_j}q'(d_i,D_j) + \frac{1}{2}\mathbb{E}(R'), \end{aligned} \tag{5.17}$$

where $q'(d_i,d_j) = q'_{ij} = I(X_i \in \Omega_k)\rho e^{-\rho\alpha_i}Q_{ij}(\beta) + I(X_j \in \Omega_k)\rho e^{-\rho\alpha_j}Q_{ji}(\beta)$, and $R' = \frac{1}{\binom{n}{2}}\sum_{i<j}q'_{ij}$. As a result, the new set of over-identifying moment conditions are $\frac{1}{n}\sum_{i=1}^{n}G_i(\beta,\rho)$, where $G_i(\beta,\rho) = (g_i^{(1)T}(\beta), g_{i1}^{(2)T}(\beta,\rho),\ldots,g_{iK}^{(2)T}(\beta,\rho))^T$. With the iteratively update algorithm (Hansen et al., 1996), the extended estimator of the true parameter (β_0,ρ_0) is obtained by minimizing the new GMM loss function:

$$\Big[\frac{1}{n}\sum_{i=1}G_i(\beta,\rho)\Big]^T W_\rho \Big[\frac{1}{n}\sum_{i=1}G_i(\beta,\rho)\Big], \tag{5.18}$$

where W_ρ is the iteratively updated weight matrix that converges in probability to the optimal weight matrix given by var$[\frac{1}{n}\sum_{i=1}G_i(\beta_0,\rho_0)]^{-1}$. Let $(\widetilde{\beta},\widetilde{\rho})$ be the extended AFT-GMM estimator, its asymptotic property is summarized in theorem 2.2.

Theorem 5.2.2. *Under regularity conditions (a) to (e), and conditions in Newey and*

McFadden (1994) Theorem 7.2, we have:

$$\sqrt{n}(\widetilde{\beta} - \beta_0) \xrightarrow{D} N(0, \Gamma_\rho^{-1}),$$

where $\Gamma_\rho^{-1} = (D_G \Sigma^{-1} D_G + B' A'^{-1} B')^{-1}$*, provided* D_G *is non-singular, where* A' *and* B' *are defined in the Appendix.*

Theorem 2.2 implies that after introducing a scaling factor ρ for the baseline hazard function, the extended AFT-GMM estimator that combines the information from both the individual-level data and the auxiliary sub-group survival information is still asymptotically more efficient than the original log-rank estimator with Gehan's weight.

5.2.3 GMM Estimator for the Cox Model

5.2.3.1 Moment Conditions

For $i = 1, 2, \ldots, n$, we assume that, given $X_i = x$, the hazard function of T follows the Cox proportional hazards (PH) model (David, 1972) $\lambda(t|x) = \lambda(t) \exp(\beta' x)$, where β is the vector of regression coefficients and $\lambda(t)$ is an unspecified baseline hazard function. Next, we re-define a few key terms in the context of the Cox PH model. First, we define $N_i(t) = I(Y_i \leq t, \Delta_i = 1)$ as the number of observed failure event prior to time t. Then we define $S^{(k)}(t, \beta) = n^{-1} \sum_{j=1}^{n} I(Y_j \geq t) \exp(\beta' X_j) X_j^{\otimes k}$,

$k = 0, 1, 2$, with $x^{\otimes 1} = x$ and $x^{\otimes 2} = x'x$. Also, let $s^{(k)}(t, \beta) = \mathbb{E}[S^{(k)}(t, \beta)]$. Therefore, the partial score function of the Cox PH model is

$$\Phi_{PL}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \{X_i - \frac{S^{(1)}(Y_i, \beta)}{S^{(0)}(Y_i, \beta)}\}. \tag{5.19}$$

Applying the functional delta method, we can show that the above partial score function has the following asymptotically i.i.d representation:

$$\Phi_{PL}(\beta) = \frac{1}{n}\sum_{i=1}^{n} h_i^{(1)}(\beta) + o_p(n^{-1/2}),$$

where

$$h_i^{(1)} = \int_0^{\infty} \left\{ X_i - \frac{s^{(1)}(t, \beta)}{s^{(0)}(t, \beta)} \right\} \left[dN_i(t) - e^{X_i'\beta} I(Y_i \geq t) \frac{\mathbb{E}\{dN(t)\}}{s^{(0)}(t, \beta)} \right]. \tag{5.20}$$

Note that $\mathbb{E}\{dN(t)\} = s^{(0)}(t, \beta_0) d\Lambda_0(t)$. Define

$$\Omega = \text{var}\Big[h_1^{(1)}(\beta_0)\Big] = \mathbb{E}\Big[h_i^{(1)}(\beta_0) h_i^{(1)}(\beta_0)^T\Big].$$

We can show that

$$\text{var}\Big[\sqrt{n}\Phi_{PL}(\beta_0)\Big] = -\mathbb{E}\Big[\frac{\partial \Phi_{PL}(\beta)}{\partial \beta}\Big|_{\beta=\beta_0}\Big] = -\mathbb{E}\Big[\frac{\partial h_1^{(1)}(\beta)}{\partial \beta}\Big|_{\beta=\beta_0}\Big] + o_p(1) = \Omega + o_p(1)$$

Therefore, let $\widehat{\beta}_{PL}$ be the maximum partial likelihood estimator of the Cox PH

model, David (1972) showed that

$$\sqrt{n}(\widehat{\beta}_{PL} - \beta_0) \xrightarrow{D} N(0, \Omega^{-1}).$$

Under the proportional hazards model, the moment conditions derived from the auxiliary sub-group survival information depend on the baseline cumulative hazard function only through its values at the landmark time point t^*. Therefore, by introducing an additional parameter $\alpha = \Lambda(t^*) = \int_0^\infty I(u \leq t^*)\lambda(u)du$, we can derive the individual cumulative hazard straightforwardly and connect it to the sub-group survival probability, which is the key to deriving the auxiliary moment condition. First, by applying functional delta method to the Breslow estimator of the cumulative baseline hazard function, we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\{\frac{\Delta_i I(Y_i \leq t^*)}{S^{(0)}(Y_i, \beta)} - \alpha\} = \frac{1}{n}\sum_{i=1}^{n} h_i^{(2)}(\alpha, \beta) + o_p(n^{-1/2}), \tag{5.21}$$

where

$$\begin{aligned} h_i^{(2)}(\alpha, \beta) = &\int_0^\infty \frac{I(t \leq t^*)}{s^{(0)}(t, \beta)}\left[dN_i(t) - e^{X_i'\beta} I(Y_i \geq t)\frac{\mathbb{E}\{dN(t)\}}{s^{(0)}(t, \beta)}\right] \\ &+ \int_0^\infty \frac{I(t \leq t^*)}{s^{(0)}(t, \beta)}\mathbb{E}\{dN(t)\} - \alpha. \end{aligned} \tag{5.22}$$

Then, by double expectation, the survival information at the landmark time t^* amounts

to the population moments $\mathbb{E}\{\Psi_k(X_i, \beta, \Lambda)\} = 0$, where

$$\Psi_k(X_i, \beta, \Lambda) = I(X_i \in \Omega_k)[\exp\{-\Lambda(t^*)\exp(\beta' X_i)\} - \phi_k], \quad k = 1, \ldots, K.$$

Thus, by treating α as a nuisance parameter, the set of moment conditions representing the auxiliary information are: for $k = 1, \ldots, K$,

$$\frac{1}{n}\sum_{i=1}^{n} h_{ik}^{(3)}(\alpha, \beta) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in \Omega_k)[\exp\{-\alpha\exp(\beta' X_i)\} - \phi_k]. \tag{5.23}$$

In summary, under the Cox PH model, the partial likelihood together with the auxiliary survival information are transformed into a set of over-identifying moment conditions, $\frac{1}{n}\sum_{i=1}^{n} H_i(\beta, \alpha)$, where $H_i(\alpha, \beta) = (h_i^{(1)T}(\beta), h_i^{(2)T}(\alpha, \beta), h_{i1}^{(3)T}(\alpha, \beta), \ldots, h_{iK}^{(3)T}(\alpha, \beta))^T$

5.2.3.2 Estimation and Inference

Applying the iterative update algorithm of GMM, we can estimate the true parameter value (α_0, β_0) by minimizing the weighted quadratic function:

$$\Big[\frac{1}{n}\sum_{i=1} H_i(\alpha, \beta)\Big]^T V \Big[\frac{1}{n}\sum_{i=1} H_i(\alpha, \beta)\Big], \tag{5.24}$$

where V is the iteratively updated weight matrix that converges to the optimal weight given by var$[\frac{1}{n}\sum_{i=1} H_i(\alpha_0, \beta_0)]^{-1}$, that is, the inverse of the covariance ma-

trix of the moment conditions evaluated at the true value. Denote the obtained minimizer as $\widehat{\beta}_c$. Its asymptotic property is summarized in the theorem 2.3.

Theorem 5.2.3. *Under the regularity conditions that X is bounded, both T and C are absolutely continuous, the domain of (α, β) is a compact set and (α_0, β_0) is in its interior, and other regularity conditions for a GMM estimator to be asymptotically Normal, we have:*

$$\sqrt{n}(\widehat{\beta}_c - \beta_0) \xrightarrow{D} N(0, \Pi^{-1}),$$

where $\Pi^{-1} = (\Omega + \widetilde{B}\widetilde{A}^{-1}\widetilde{B})^{-1}$, provided Π is non-singular, where $\widetilde{A}$ and $\widetilde{B}$ are defined in the Appendix.

As we can see, theorem 2.2 implies that the GMM estimator that combines the information from both the individual-level data and the auxiliary sub-group survival information is asymptotically more efficient than the original maximum partial likelihood estimator of the Cox PH model.

5.2.3.3 Account for the Inconsistency in Baseline Hazard

As in section 2.2.3, we assume

$$\lambda_0^a(t) = \rho\lambda_0(t), \text{ with } \rho > 0$$

Under the proportional hazards model, the auxiliary information is summarized by

$$\frac{1}{n}\sum_{i=1}^{n} h_{ik}^{(3)}(\alpha,\beta,\rho) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in \Omega_k)[\exp\{-\rho\alpha\exp(\beta' X_i)\} - \phi_k]$$

As a result, the new set of over-identifying moment conditions are $\frac{1}{n}\sum_{i=1}^{n} H_i(\alpha,\beta,\rho)$, where $H_i(\alpha,\beta,\rho) = (h_i^{(1)T}(\beta), h_i^{(2)T}(\alpha,\beta), h_{i1}^{(3)T}(\alpha,\beta,\rho), \ldots, h_{iK}^{(3)T}(\alpha,\beta,\rho))^T$.

Following the GMM estimating procedure, we can estimate the true parameter value $(\alpha_0,\beta_0,\rho_0)$ by minimizing the weighted quadratic function:

$$\Big[\frac{1}{n}\sum_{i=1} H_i(\alpha,\beta,\rho)\Big]^T V_\rho \Big[\frac{1}{n}\sum_{i=1} H_i(\alpha,\beta,\rho)\Big]$$

where V_ρ is the iteratively updated weight matrix that converges to the optimal weight given by var$[\frac{1}{n}\sum_{i=1} H_i(\alpha_0,\beta_0,\rho_0)]^{-1}$. Let $\widetilde{\beta}_c$ be the obtained GMM estimate. Its asymptotic property is summarized in theorem 2.4.

Theorem 5.2.4. *Under the regularity conditions that X is bounded, both T and C are absolutely continuous, the space of (α,β,ρ) is a compact set with $(\alpha_0,\beta_0,\rho_0)$ is in its interior, and other regularity conditions for a GMM estimator to be asymptotically Normal, we have:*

$$\sqrt{n}(\widetilde{\beta}_c - \beta_0) \xrightarrow{D} N(0, \Pi_\rho^{-1})$$

where $\Pi_\rho^{-1} = (\Omega + \dot{B}\dot{A}^{-1}\dot{B})^{-1}$*, provided* Π_ρ *is non-singular, where* $\dot{A}$ *and* $\dot{B}$ *are defined in the Appendix.*

5.3 Simulation Studies

5.3.1 Data Simulation

To evaluate the finite-sample performance of the proposed GMM estimators, we conducted two sets of simulation studies. In all simulations, the covariate X_1 was simulated from a standard normal distribution, and X_2 was simulated independently of X_1 from a Bernoulli distribution with mean 0.5. Let $X_3 = X_1X_2$ be the interaction term of X_1 and X_2. The failure time T was simulated from a Weibull distribution with scale parameter $\exp(0.5 \times X_1 - 0.5 \times X_2 + 0.5 \times X_3)$ and shape parameter 2. This data generation mechanism satisfies both the AFT model and the Cox model, with the baseline hazard function $\lambda_0(t) = 2t$. Under the AFT model, the true value of regression coefficients are $\beta_1^* = 0.5, \beta_2^* = -0.5, \beta_3^* = 0.5$. Under the Cox PH model, the true regression parameter values are $\beta_1^* = -1, \beta_2^* = 1, \beta_3^* = -1$. The censoring time C was generated from a gamma distribution, whose parameters are selected so that the censoring rate was approximately 0%, 30% or 50%. The sub-group membership is defined as $\Omega_1 = \{(X_1, X_2) : X_1 \leq 0, X_2 = 0\}$ and $\Omega_2 = \{(X_1, X_2) : X_1 > 0, X_2 = 0\}$. This setting aims to mimic the situation of

a randomized clinical trial where X_1 is a continuous baseline risk factor, such as age, X_2 is the treatment/control assignment indicator, and only the untreated group ($X_2 = 0$) has auxiliary sub-group survival information. In each simulation, 500 datasets were generated independently, each with a sample size of $n = 150$.

In the first set of simulations, we consider the scenario where the baseline hazard function for the individual-level data is consistent with the auxiliary aggregated data, that is, $\lambda_0^a(t) = \lambda_0(t)$. As a result, at the landmark time points $t^* = (0.4, 0.8, 1.2)$, the corresponding auxiliary survival probabilities were $(0.764, 0.396, 0.162)$ for subjects in Ω_1, and $(0.946, 0.803, 0.620)$ for subjects in Ω_2. In the second set of simulations, we set $\lambda_0^a(t) = 1.5\lambda_0(t)$, that is, $\rho_0 = 1.5$. Therefore, at the same landmark time points $t^* = (0.4, 0.8, 1.2)$, the survival rates are $(0.673, 0.270, 0.078)$ for Ω_1, and $(0.921, 0.724, 0.500)$ for Ω_2.

5.3.2 Results of the AFT-GMM Estimator

Table 5.1 and 5.2 summarize the empirical bias and empirical standard error of the proposed AFT-GMM estimators in the two simulation scenarios respectively. The relative efficiency of the two AFT-GMM estimators comparing to the benchmark log-rank estimator with Gehan-type weights was also reported in the table. In each table, the first column shows the expected proportion of censoring. The second column shows the number of auxiliary survival rates that are utilized by the estimator per subgroup. For example, Num. t^* = 1 means that only the survival

probabilities at $t^* = 0.4$ are used, and Num. t^* = 2 indicates that only the information at $t^* = 0.4$ and $t^* = 0.8$ are used. $\beta_1, \beta_2, \beta_3$ are the regression coefficients under the accelerated failure time assumption and ρ is the scaling factor for the baseline hazard function in the extended estimator.

Shown in Table 5.1, in the situation where the auxiliary survival information is consistent with the individual-level data, all the bias estimates are roughly between -0.01 and 0.01, which suggests that with or without the scaling factor, the AFT-GMM estimators are empirically unbiased. Moreover, both GMM estimators are more efficient in terms of mean square error than the log-rank AFT estimator. The efficiency gain with respect to β_1 is quite significant in both GMM estimators, because X_1 is the primary factor of the sub-group definition. The relative efficiency in estimating β_2 and β_3 are relatively smaller, around 1.1 to 1.5. Also, for both GMM estimators, the relative efficiency tends to increase with the proportion of censoring, at least within the tested range from 0% to 50%. When 50% of the observations are censored, the GMM estimators can be at most 5 times more efficient than the log-rank estimator. Besides, incorporating survival information at two time points is uniformly more efficient than using only one time points. But utilizing all three time points is not necessarily more efficient than using only two, especially when the censoring proportion is low. A possible explanation is that the auxiliary moments within the same sub-group are highly correlated. Intuitively, it makes the incremental information per additional time point drops quickly as more time points

are used. Comparing to the base version, the extended AFT-GMM estimator is basically as efficient in estimating β_1 and β_3 but less efficient for β_2, because an additional scaling parameter ρ also has to be estimated.

In Table 5.2, however, when the auxiliary aggregate data has a scaled-down baseline hazard function comparing to the individual-level data, the AFT-GMM estimator without accounting for inconsistent hazard function is biased in estimating β_2 at a scale of 20% to 30%. The extended estimator, on the other hand, still provides empirically unbiased estimates to all the regression coefficients. It is also more efficient than the log-rank AFT estimator, and the relative efficiency on each parameter is close to what we observe in Table 5.1 when the information is consistent. Figure 5.1 and 5.2 show the density curves of the empirical sampling distributions of the two AFT-GMM estimators and the log-rank estimator. In summary, with finite sample size, when the auxiliary information is consistent with individual-level data, both GMM estimators are more efficient than the log-rank estimator and the base version is slightly bettern than the other; when the information is not consistent, the AFT-GMM estimator without adjusting for inconsistent baseline hazard function is biased, while the extended estimator is still unbiased and more efficient than then benchmark.

Table 5.1: Summary Statistics of the AFT-GMM Estimators given Consistent Auxiliary Information

Pr(Cens)	Num. t*	Method	β_1			β_2			β_3			ρ	
			Bias	SE	RE	Bias	SE	RE	Bias	SE	RE	Bias	SE
0%	1	GMM	0.005	0.054	1.670	0.011	0.090	1.080	-0.011	0.087	1.080		
		GMM$_\rho$	0.000	0.056	1.560	0.009	0.095	0.980	-0.004	0.085	1.150	0.058	0.248
	2	GMM	0.008	0.042	2.680	0.006	0.078	1.480	-0.013	0.083	1.340		
		GMM$_\rho$	0.006	0.042	2.720	0.003	0.092	1.070	-0.010	0.083	1.350	0.035	0.144
	3	GMM	0.005	0.045	2.330	0.003	0.082	1.230	-0.005	0.087	1.240		
		GMM$_\rho$	0.005	0.044	2.440	0.005	0.092	0.980	-0.005	0.090	1.160	0.007	0.126
30%	1	GMM	0.004	0.061	1.930	0.006	0.115	1.070	0.004	0.113	1.230		
		GMM$_\rho$	-0.002	0.064	1.760	0.006	0.120	0.980	0.009	0.113	1.230	0.060	0.266
	2	GMM	0.006	0.050	3.180	0.006	0.097	1.550	-0.001	0.111	1.370		
		GMM$_\rho$	0.005	0.046	3.770	0.005	0.114	1.130	0.001	0.110	1.400	0.024	0.165
	3	GMM	0.011	0.054	3.070	0.003	0.102	1.450	-0.009	0.120	1.170		
		GMM$_\rho$	0.011	0.052	3.300	-0.011	0.119	1.060	-0.008	0.114	1.300	0.044	0.165
50%	1	GMM	0.017	0.087	1.880	0.000	0.158	1.010	0.003	0.161	1.100		
		GMM$_\rho$	0.005	0.081	2.240	0.002	0.161	0.980	0.008	0.149	1.280	0.087	0.288
	2	GMM	0.011	0.056	4.430	0.012	0.124	1.450	-0.005	0.140	1.350		
		GMM$_\rho$	0.012	0.052	5.060	0.003	0.148	1.030	-0.006	0.139	1.370	0.047	0.209
	3	GMM	0.014	0.057	4.620	0.009	0.132	1.520	-0.011	0.142	1.300		
		GMM$_\rho$	0.013	0.052	5.540	0.001	0.157	1.080	-0.008	0.140	1.340	0.056	0.218

β_1, β_2 and β_3 are the regression coefficients of interest. ρ is the scaling factor in the extended GMM estimator. Pr(Cens) shows the proportion of censoring. Num. t* is the number of landmark time utilized per group. In the Method column, GMM represents the unadjusted AFT-GMM estimator, and GMMρ stands for the extended estimator with the baseline hazard adjusted by ρ. Bias, SE, RE are the empirical bias, empirical standard error and empirical relative efficiency.

Table 5.2: Summary Statistics of the AFT-GMM Estimators given Inconsistent Auxiliary Information

Pr(Cens)	Num. t*	Method	β_1			β_2			β_3			ρ	
			Bias	SE	RE	Bias	SE	RE	Bias	SE	RE	Bias	SE
0%	1	GMM	0.043	0.060	0.900	0.087	0.100	0.500	-0.045	0.092	0.790		
		GMM$_\rho$	0.004	0.057	1.500	0.015	0.099	0.900	-0.009	0.085	1.140	0.070	0.383
	2	GMM	0.012	0.054	1.600	0.142	0.097	0.310	-0.014	0.095	1.020		
		GMM$_\rho$	0.009	0.043	2.540	0.019	0.103	0.850	-0.013	0.083	1.340	0.002	0.227
	3	GMM	0.004	0.054	1.630	0.137	0.099	0.290	-0.005	0.100	0.940		
		GMM$_\rho$	0.008	0.044	2.390	0.038	0.109	0.700	-0.008	0.096	1.020	-0.055	0.214
30%	1	GMM	0.046	0.071	1.010	0.100	0.131	0.520	-0.028	0.125	0.960		
		GMM$_\rho$	0.004	0.067	1.600	0.014	0.125	0.910	0.003	0.119	1.110	0.069	0.397
	2	GMM	0.001	0.055	2.670	0.159	0.113	0.380	0.000	0.123	1.120		
		GMM$_\rho$	0.007	0.046	3.730	0.031	0.128	0.890	0.000	0.109	1.420	-0.028	0.272
	3	GMM	0.003	0.059	2.670	0.147	0.123	0.410	-0.015	0.128	1.020		
		GMM$_\rho$	0.009	0.051	3.470	0.035	0.140	0.770	-0.003	0.121	1.150	-0.021	0.276
50%	1	GMM	0.055	0.103	1.080	0.120	0.170	0.580	-0.012	0.172	0.960		
		GMM$_\rho$	0.012	0.086	1.960	0.018	0.163	0.950	0.004	0.151	1.250	0.089	0.428
	2	GMM	0.004	0.064	3.510	0.165	0.142	0.470	-0.001	0.152	1.150		
		GMM$_\rho$	0.015	0.051	5.100	0.041	0.159	0.890	-0.007	0.138	1.390	-0.014	0.323
	3	GMM	-0.002	0.059	4.570	0.148	0.146	0.620	-0.015	0.152	1.130		
		GMM$_\rho$	0.010	0.052	5.680	0.051	0.181	0.810	-0.008	0.155	1.090	-0.024	0.370

β_1, β_2 and β_3 are the regression coefficients of interest. ρ is the scaling factor in the extended GMM estimator. Pr(Cens) shows the proportion of censoring. Num. t* is the number of landmark time utilized per group. In the Method column, GMM represents the unadjusted AFT-GMM estimator, and GMMρ stands for the extended estimator with the baseline hazard adjusted by ρ. Bias, SE, RE are the empirical bias, empirical standard error and empirical relative efficiency.

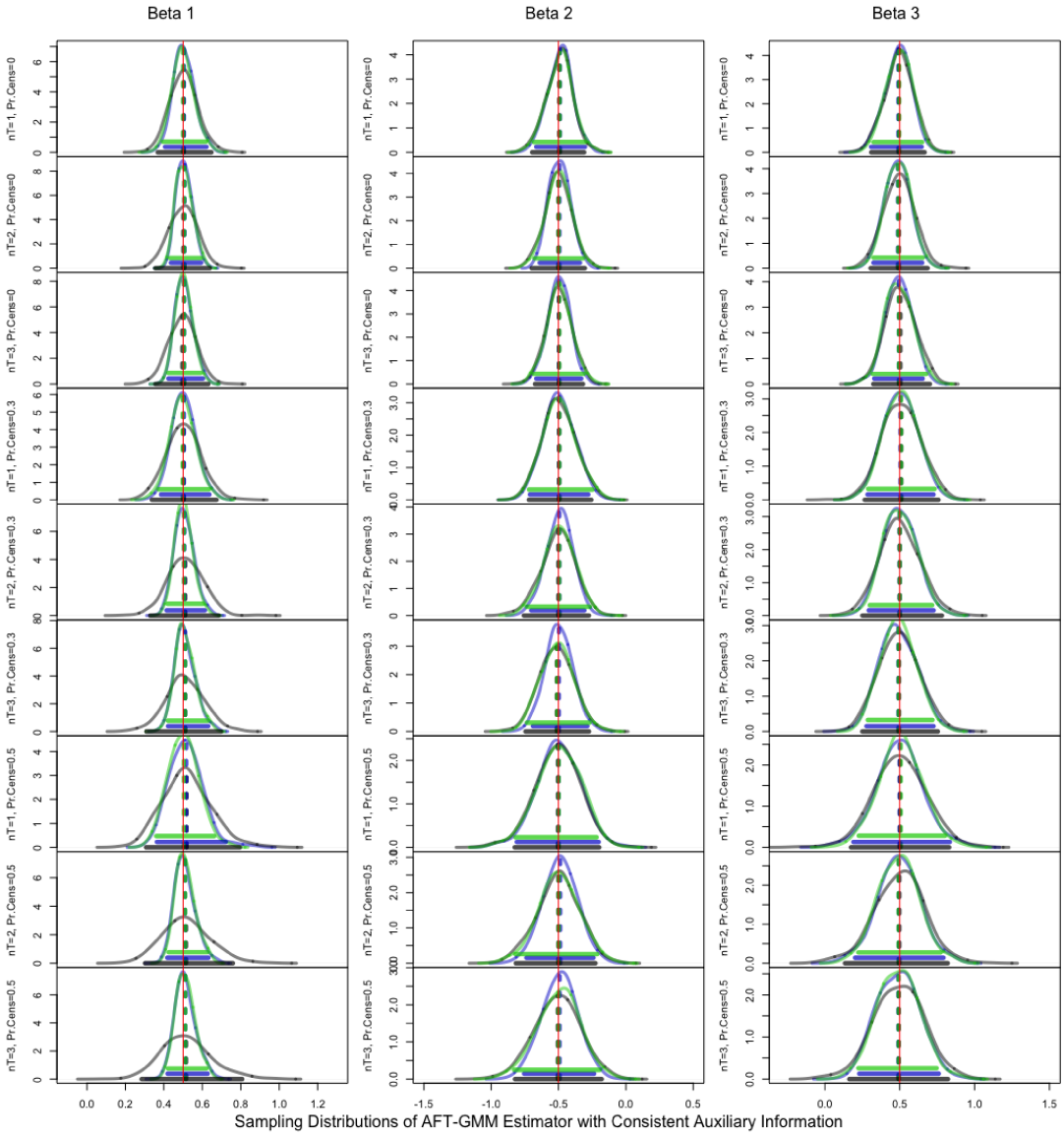
Beta 1
Beta 2
Beta 3
nT=1, Pr.Cens=0
nT=2, Pr.Cens=0
nT=3, Pr.Cens=0
nT=1, Pr.Cens=0.3
nT=2, Pr.Cens=0.3
nT=3, Pr.Cens=0.3
nT=1, Pr.Cens=0.5
nT=2, Pr.Cens=0.5
nT=3, Pr.Cens=0.5
Sampling Distributions of AFT-GMM Estimator with Consistent Auxiliary Information

Fig. 5.1: The density curves represent the empirical sampling distributions of the two AFT-GMM estimators and the log-rank estimator given the auxiliary survial information is **consistent** with the individual-level data. Three columns of plots correspond to the three estimtated regression coefficients respectively. Each row represents a unique model fitting configuration as labeled on the y-axis where nT stands for the number of survival probabilities utilized per group, and Pr.Cens is the expected censoring probability. In each individual plot, gey color corresponds to the log-rank estimator, blue color represents the unadjusted GMM estimator, and gree color is the extended GMM estimator. The horizontal bars stand for the $(2.5\%, 97.5\%)$ intervals of the corresponding curves. The vertical dashed lines are the means of the sampling distributions. The vertical red lines mark the true value of the parameters.

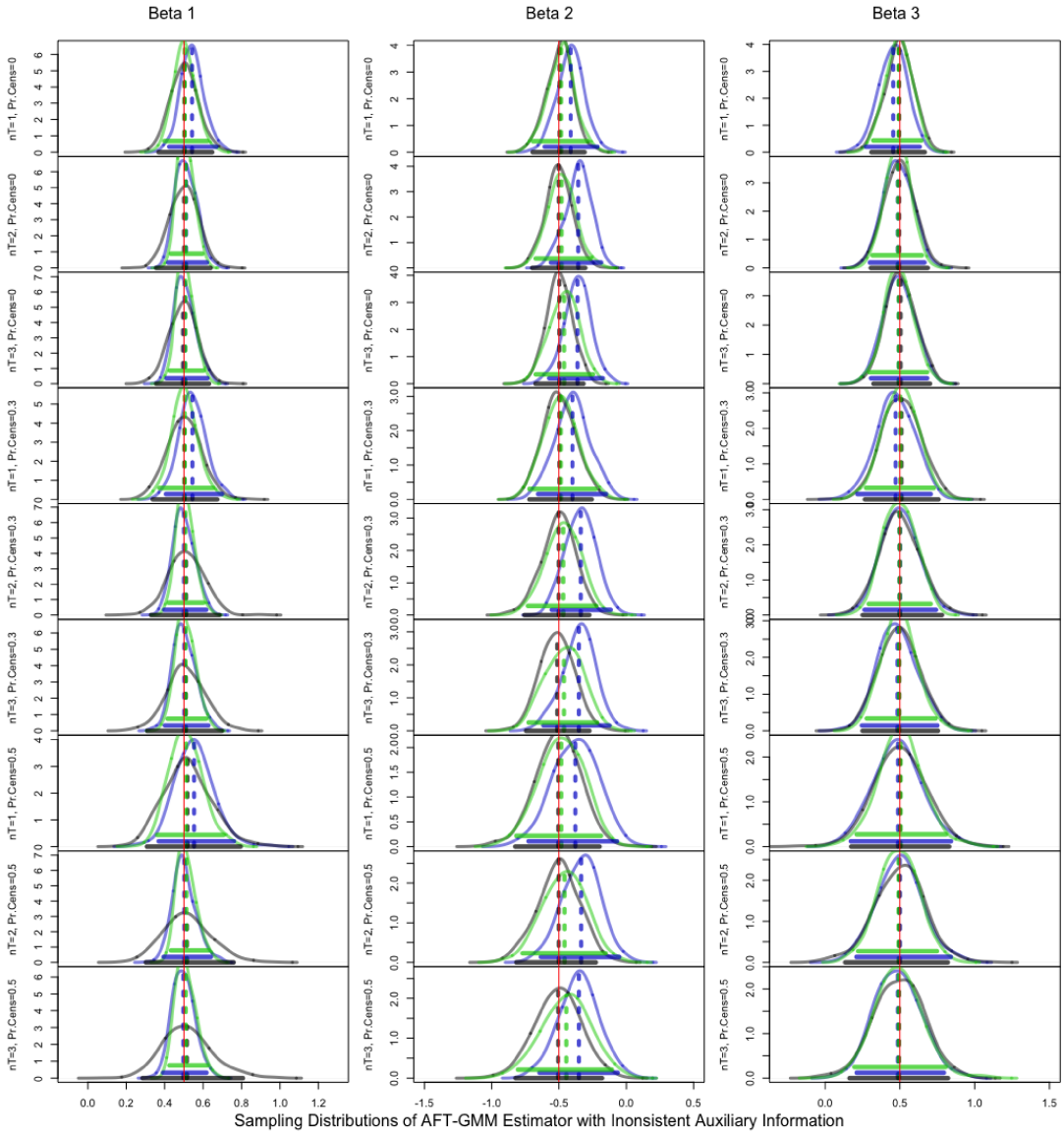

Fig. 5.2: The density curves represent the empirical sampling distributions of the two AFT-GMM estimators and the log-rank estimator given the auxiliary survial information is **inconsistent** with the individual-level data. The graph legends are the same as in figure 5.1.

5.3.3 Results of the Cox-GMM Estimator

Table 5.3 and 5.4 summarize the empirical bias, empirical standard error, and the relative efficiency of the proposed Cox-GMM estimators in the two simulation scenarios respectively. Similar to Table 5.1 in the above section, the first column lists the expected proportion of censoring. The second column lists the number of auxiliary survival rates that are utilized by the estimator per subgroup. $\beta_1, \beta_2, \beta_3$ are the regression coefficients under the proportional hazards assumption and ρ is the scaling factor for the baseline hazard function in the extended Cox-GMM estimator.

Table 5.3 shows that, when the auxiliary survival information is consistent with the individual-level data, the empirical bias of $\beta_1, \beta_2, \beta_3$ for both Cox-GMM estimators are very small. But the extended Cox-GMM estimator tends to overestimate ρ by about 10% when the censoring proportion is large. Moreover, both Cox-GMM estimators are more efficient than the partial-likelihood estimator. The largest efficiency gain, for both estimators, is obtained for estimating β_1, again because X_1 is the primary factor of the sub-group definition. The relative efficiency in estimating β_2 and β_3 are generally below 2. Also, the efficiency gain tends to increase with the proportion of censoring, at least within the tested range from 0% to 50%. When 50% of the observations are censored, the GMM estimator can be at most 5 times more efficient than the partial-likelihood estimator. But unlike the AFT-GMM estimators which are generally more efficient when using more than one time point information, the Cox-GMM estimators of β_1 are most efficient when utilizing only

one time point information. The extended Cox-GMM estimator is as efficient as the original Cox-GMM estimator for β_1 and β_3, but has no efficiency gain for β_2. In Table 5.4, when the auxiliary information is inconsistent, the base version Cox-GMM estimator underestimates the value of β_2 by a scale of 20% to 30%. The extended estimator, on the other hand, still provides empirically unbiased estimates to all the regression coefficients, and it is is still more efficient than the paritial-likelihood estimator. The percentage of efficiency gain on each parameter is close to what we observe when the information is consistent. Figure 5.3 and 5.4 provide a comprehensive visualization of the empirical sampling distributions of the two Cox-GMM estimators and the partial-likelihood estimator. In summary, with finite sample size, when the auxiliary information is consistent with individual-level data, both Cox-GMM estimators are more efficient than the partial-likelihood estimator; when the information is not consistent, the original Cox-GMM estimator is biased and less efficient than the extended model which is still unbiased.

Table 5.3: Summary Statistics of the Cox-GMM Estimators given Consistent Auxiliary Information

Pr(Cens)	Num. t*	Method	β_1			β_2			β_3			ρ	
			Bias	SE	RE	Bias	SE	RE	Bias	SE	RE	Bias	SE
0%	1	GMM	-0.014	0.080	3.57	-0.004	0.154	1.41	0.026	0.170	1.21	0.058	0.269
		GMM$_\rho$	-0.013	0.079	3.67	0.016	0.185	0.98	0.008	0.177	1.14		
	2	GMM	-0.016	0.083	3.06	-0.003	0.146	1.66	0.002	0.171	1.23	0.084	0.207
		GMM$_\rho$	-0.020	0.085	2.87	0.038	0.188	1.00	-0.014	0.170	1.24		
	3	GMM	-0.020	0.088	2.78	0.001	0.149	1.60	0.013	0.183	1.22	0.065	0.206
		GMM$_\rho$	-0.024	0.091	2.56	0.037	0.190	0.98	0.004	0.182	1.23		
30%	1	GMM	-0.023	0.092	4.15	-0.005	0.190	1.45	0.022	0.222	1.44	0.082	0.314
		GMM$_\rho$	-0.022	0.092	4.17	0.026	0.220	1.08	-0.002	0.224	1.43		
	2	GMM	-0.020	0.097	3.33	-0.019	0.182	1.62	0.001	0.213	1.62	0.104	0.250
		GMM$_\rho$	-0.028	0.092	3.53	0.035	0.231	1.02	-0.013	0.218	1.54		
	3	GMM	-0.021	0.102	2.96	-0.012	0.172	1.83	0.006	0.245	1.13	0.078	0.216
		GMM$_\rho$	-0.026	0.101	2.96	0.029	0.236	0.98	0.002	0.228	1.30		
50%	1	GMM	-0.023	0.094	5.39	-0.012	0.230	1.61	0.014	0.263	1.49	0.101	0.362
		GMM$_\rho$	-0.023	0.092	5.62	0.025	0.285	1.05	-0.009	0.268	1.44		
	2	GMM	-0.019	0.091	5.70	-0.027	0.217	1.69	-0.022	0.254	1.52	0.123	0.279
		GMM$_\rho$	-0.028	0.094	5.12	0.040	0.272	1.09	-0.026	0.254	1.52		
	3	GMM	-0.022	0.099	4.92	-0.035	0.211	1.97	-0.010	0.271	1.39	0.136	0.287
		GMM$_\rho$	-0.030	0.099	4.73	0.039	0.289	1.08	-0.005	0.259	1.52		

β_1, β_2 and β_3 are the regression coefficients of interest. ρ is the scaling factor in the extended GMM estimator. Pr(Cens) shows the proportion of censoring. Num. t* is the number of landmark time utilized per group. In the Method column, GMM represents the unadjusted Cox-GMM estimator, and GMMρ stands for the extended estimator with the baseline hazard adjusted by ρ. Bias, SE, RE are the empirical bias, empirical standard error and empirical relative efficiency.

Table 5.4: Summary Statistics of the Cox-GMM Estimators given Inconsistent Auxiliary Information

Pr(Cens)	Num. t*	Method	β_1			β_2			β_3			ρ	
			Bias	SE	RE	Bias	SE	RE	Bias	SE	RE	Bias	SE
0%	1	GMM	0.018	0.081	3.42	-0.199	0.155	0.53	0.130	0.171	0.77	0.095	0.404
		GMM$_\rho$	-0.017	0.079	3.60	0.017	0.185	0.98	0.010	0.177	1.14		
	2	GMM	0.010	0.093	2.50	-0.239	0.145	0.45	0.082	0.179	0.93	0.118	0.307
		GMM$_\rho$	-0.023	0.085	2.82	0.039	0.188	1.00	-0.012	0.169	1.25		
	3	GMM	-0.001	0.106	2.01	-0.250	0.509	0.11	0.072	0.202	0.89	0.085	0.283
		GMM$_\rho$	-0.027	0.087	2.73	0.032	0.185	1.04	0.009	0.179	1.27		
30%	1	GMM	0.011	0.092	4.35	-0.226	0.190	0.60	0.113	0.224	1.14	0.133	0.471
		GMM$_\rho$	-0.027	0.093	3.98	0.028	0.220	1.08	0.001	0.224	1.43		
	2	GMM	0.013	0.105	2.92	-0.287	0.180	0.47	0.044	0.212	1.56	0.145	0.371
		GMM$_\rho$	-0.031	0.094	3.33	0.036	0.232	1.01	-0.010	0.219	1.53		
	3	GMM	0.005	0.115	2.43	-0.295	0.215	0.41	0.032	0.258	1.00	0.098	0.310
		GMM$_\rho$	-0.028	0.103	2.82	0.028	0.228	1.05	0.006	0.234	1.23		
50%	1	GMM	0.014	0.094	5.59	-0.266	0.227	0.70	0.075	0.260	1.42	0.160	0.546
		GMM$_\rho$	-0.028	0.093	5.36	0.027	0.285	1.05	-0.005	0.268	1.44		
	2	GMM	0.011	0.104	4.50	-0.321	0.218	0.54	-0.012	0.264	1.42	0.177	0.416
		GMM$_\rho$	-0.032	0.095	4.90	0.041	0.272	1.09	-0.022	0.254	1.52		
	3	GMM	0.002	0.108	4.33	-0.323	0.220	0.59	-0.003	0.272	1.38	0.190	0.437
		GMM$_\rho$	-0.032	0.101	4.51	0.039	0.292	1.06	0.000	0.262	1.49		

β_1, β_2 and β_3 are the regression coefficients of interest. ρ is the scaling factor in the extended GMM estimator. Pr(Cens) shows the proportion of censoring. Num. t* is the number of landmark time utilized per group. In the Method column, GMM represents the unadjusted Cox-GMM estimator, and GMMρ stands for the extended estimator with the baseline hazard adjusted by ρ. Bias, SE, RE are the empirical bias, empirical standard error and empirical relative efficiency.

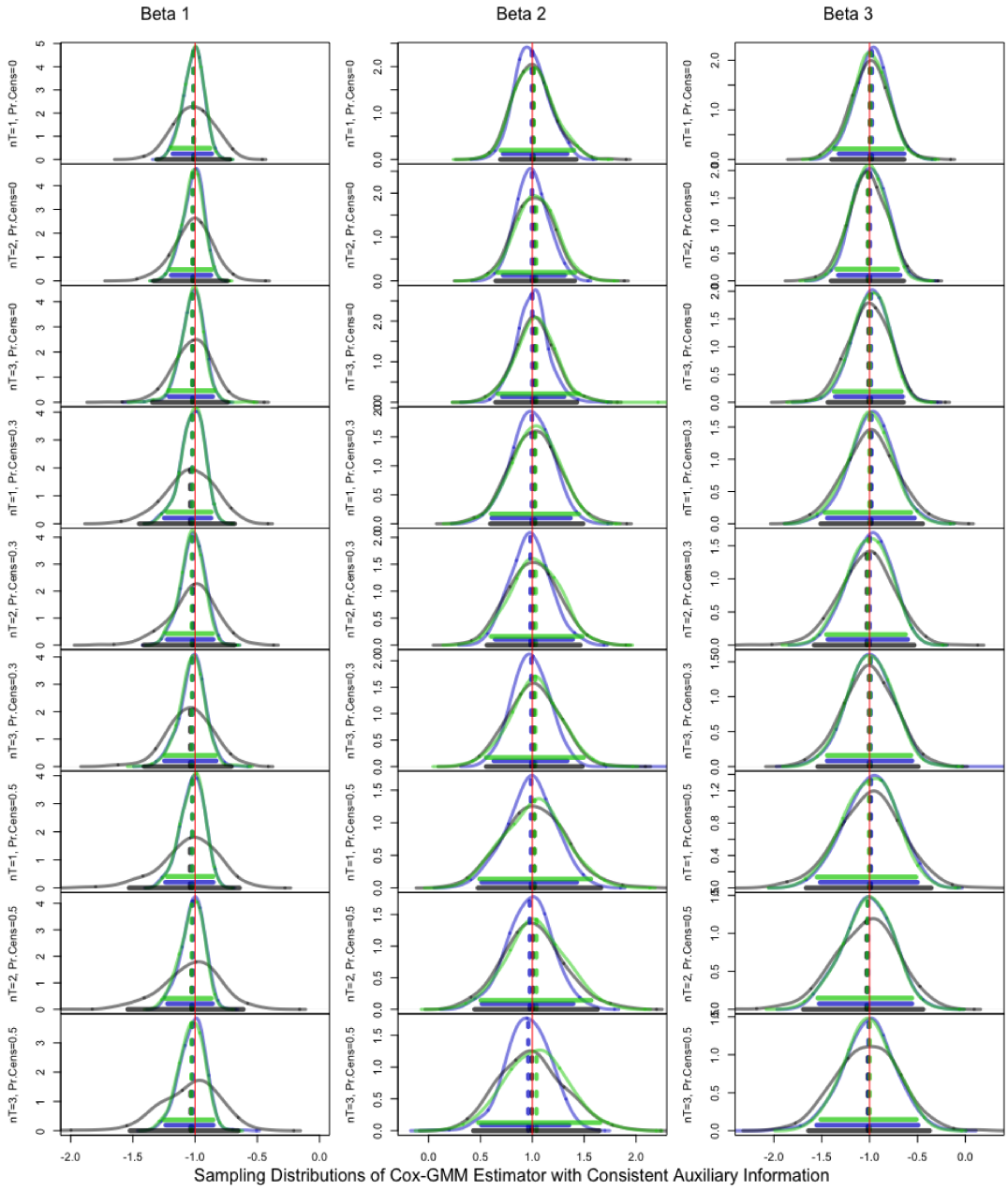

Fig. 5.3: The density curves represent the empirical sampling distributions of the two Cox-GMM estimators and the partial-likelihood estimator given the auxiliary survial information is **consistent** with the individual-level data. The graph legends are the same as in figure 5.1

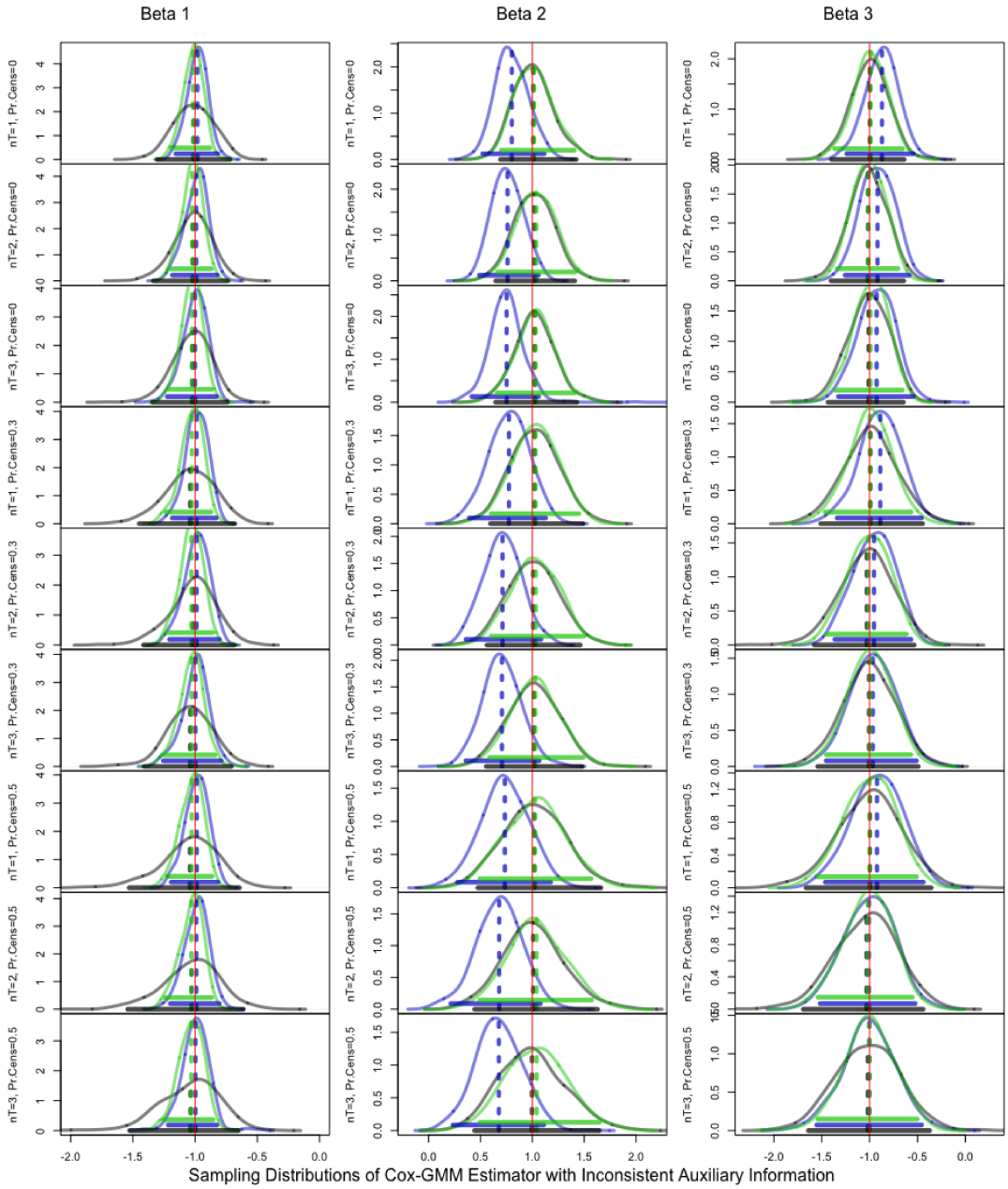

Fig. 5.4: The density curves represent the empirical sampling distributions of the two Cox-GMM estimators and the partial-likelihood estimator given that the auxiliary survial information is **inconsistent** with the individual-level data. The graph legends are the same as in figure 5.1.

5.4 Pancreatic Cancer Prognosis Analysis

Pancreatic ductal adenocarcinoma (PDAC) (Stark and Eibl, 2015) is the most common form of pancreatic malignancy. Despite the advancement in the knowledge of tumor biology and the improvement in diagnosis and health care, the prognosis remains strikingly poor. Radical surgical resection is so far the only clinically beneficial treatment for PDAC in terms of overall survival. However, at the time of diagnosis, no more than 20% of patients with PDAC have surgically resectable condition. Furthermore, pancreatic cancer may recur within 5 years for most patients.

We apply the proposed GMM estimators to analyze the data from a recent retrospective cohort study in order to quantify the impact of risk factors on patient survival following pancreatectomy. This study collected data from 209 consecutive patients who had surgical resection of PDAC and follow-up at the Johns Hopkins Hospital from Jan 9, 1998 to Jun 13, 2007. The dataset includes the patients demographic information and lab test results, clinical and pathological exam results, treatment data, and dates of death, including all-cause and cancer-specific deaths. Previous researches indicate that the most important prognostic factors for survival after pancreatectomy are tumor characteristics, among which negative resection margin, negative lymph node, and absence of perineural invasion are favorable. Thus, we included the following covariates in our regression analysis: presence of

lymph nodes (Node), positive resection margins (Margin), presence of perineural invasion (PNI), age group (> 65 or not) at surgery, and gender. Useful auxiliary survival information for PDAC is found in Cameron et al. (2006), where the $3-$year survival probabilities are reported for four patient sub-groups: (1) node-negative: $\phi_1 = 0.40$, (2) node-positive: $\phi_2 = 0.26$, (3) margin-negative: $\phi_3 = 0.35$, and (4) margin-positive: $\phi_4 = 0.20$. These probabilities were estimated based on 1000 consecutive pancreatectomies performed by a single surgeon from March 1969 to May 2003.

At the model fitting stage, we first applied the base version of the AFT-GMM and Cox-GMM estimators, and it turned out that the Sargan-Hansen J-test rejected the null with 95% confidence in both cases. This is as expected since the auxiliary information source is not a population level database, which indicates that the baseline hazard function in the auxiliary aggregate data could be different from our data at hand. Afterward, we fitted the regression using the extended GMM estimators, and the J-test did not reject the null. Table 5.5 lists the coefficient and bootstrap standard error estimates returned by the extended AFT-GMM estimator and by the log-rank AFT estimator with Gehan' weight. The two sets of coefficient estimates are almost identical, except that the estimated association parameter for the presence of lymph nodes by AFT-GMM estimator is less negative than that by the log-rank estimator. The estimated value of ρ is 0.80, which deviates from 1 with statistical significance. It implies that subjects reported by Cameron et al. (2006) have lower

baseline risks than patients in our clinical study do. This finding is also consistent with Huang et al. (2016). As for the standard error estimates, although we can estimate the asymptotic variance by the perturbation algorithm, given the relatively small sample size, we adopted the non-parametric boostrapping technique to obtain the standard errors for the estimated regression coefficients. Specifically, the 209 subjects were sampled with replacements for 1000 times independently, then the standard errors were estimated by the standard deviation of the 1000 estimates. These bootstrap standard errors in Table 5.5 display an expected pattern: the AFT-GMM estimator provides apparently smaller standard errors for the group membership related variables, node and margin status. The standard errors of other coefficients are basically the same across the two estimators. Suppose that the AFT assumption and the mean specification are correct so that these estimates are unbiased, the relative efficiency for node and margin status are about 3.61 and 2.09 respectively. As we can see, a direct benefit that our AFT-GMM estimator brings to the study is that with the increased precision, the coefficient for node status now can be declared as statistically significant, while we cannot draw this conclusion with the study sample alone.

Table 5.6 lists the coefficient and bootstrap standard error estimates returned by the extended Cox-GMM estimator and by the partial-likelihood estimator. The two sets of coefficient estimates are again almost identical, except that the estimated association parameter for node status by Cox-GMM estimator is less positive than

Table 5.5: Parameter Estimation of the AFT Model for the Pancreatic Cancer Study

	Node		Margin		PNI		>65 years		Male		ρ	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Gehan	-0.35	0.19	-0.43	0.13	-1.20	0.44	-0.21	0.12	0.17	0.13		
GMM$_\rho$	-0.29	0.10	-0.44	0.09	-1.22	0.44	-0.21	0.12	0.17	0.13	0.80	0.07

Coef and SE represent the estimated coefficient and the bootstrap standard error respectively.Gehan stands for the log-rank estimator of AFT model with Gehan weights. GMM$_\rho$ stands for the extended AFT-GMM estimator with baseline hazard adjusted by scaling factor ρ.

that by the partial-likelihood estimator. The signs of these coefficients are exactly opposite to those in Table 5.6. This implies that the AFT models and the Cox models are showing associations in the same direction. Also, the estimated value of ρ is 0.79, which is identical to the result of AFT-GMM and to that in Huang et al. (2016). The bootstrap standard errors in Table 5.6 show the same pattern as in Table 5.5. Suppose that the proportional hazards assumption and the mean specification are correct so that these estimates are unbiased, the relative efficiency for node and margin status are about 5.44 and 8.03 respectively.

Table 5.6: Parameter Estimation of the Cox Model for the Pancreatic Cancer Study

	Node		Margin		PNI		>65 years		Male		ρ	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
PL	0.37	0.21	0.41	0.17	1.09	0.38	0.28	0.16	-0.29	0.15		
GMM$_\rho$	0.28	0.09	0.36	0.06	1.12	0.37	0.28	0.15	-0.28	0.14	0.79	0.08

Coef and SE represent the estimated coefficient and the bootstrap standard error respectively.PL stands for the partial-likelihood estimator of the Cox PH model. GMM$_\rho$ stands for the extended Cox-GMM estimator with baseline hazard adjusted by scaling factor ρ.

5.5 Discussion

In this work, we proposed four GMM-based semi-parametric regression approaches for combining information from both individual-level survival data with right-censoring and the auxiliary survival information. The first two approaches are developed under the accelerated failure time assumption with the log-rank estimator as the benchmark, and the second two are developed under the proportional hazards assumption with the partial-likelihood estimator as the benchmark. Under each model assumption, one estimator is developed for getting the most efficiency gain in the situation where uses are sure that the auxiliary information is consistent with the study data, and the other estimator is built with the capacity to adjust for potential inconsistency in the baseline hazard function. We proved that under mild regularity conditions, our estimators are asymptotically more efficient than the benchmark estimators. We also demonstrated by simulation studies that the base version estimators attain the most efficiency gain provided that the auxiliary information is consistent with the study data, and the extended estimators are empirically more efficient with or without consistent information on the baseline hazard function.

By proposing these estimators, we provide two primary contributions. First, we introduce a novel approach for information synthesis using the GMM framework. The difference between our approach and the previous ones is that we do

not seek to re-invent the wheel. Basically, we do not seek to propose a new optimization procedure and work out the asymptotic properties from scratch. Instead, by transforming the moment conditions to proper asymptotic forms, the state-of-art GMM estimating procedure and properties can be applied readily. Second, our AFT-GMM approaches fill the gap of utilizing auxiliary information under the accelerated failure time assumption. Results show the efficiency improvement is significant especially when the proportion of censoring is large or when auxiliary information is available for multiple time points. As the AFT model is one of the two most widely used regression models for right-censored survival data, our method is destined to have a significant impact on the practice of time-to-event analysis. In a addtion, our Cox-GMM aproach is the GMM solution to the asymptotically equivalent problem that was previously solved by the emprirical likelihood method proposed by Huang et al. (2016). Although the Cox-GMM estimator is asymptotically at most as efficient as the EL approach, it offers better computational stability for finite sample applications.

Chapter 6

Discussion and Future Work

In this thesis, we have presented statistical methods in support of the goal of individualized health in two parts. In Part I, the Bayesian latent sparse correlation model (BLSCM) is developed to estimate the latent etiology distribution given imperfect measurements. In BLSCM, the latent etiologic state is parameterized as a multivariate binary vector, where each indicator represents the presence/absence of an etiologic agent. The measurements are assumed to be conditionally independent given the latent state, governed by the true positive rates and false positive rates. Experts' knowledge on the competition mechanism among etiologic agents is translated into a sparse correlation structure of the latent state. A scalable MALA-within-Gibbs sampling algorithm with pseudo-likelihood is proposed for approximating the exact posterior distribution. Also, a variational Bayesian algorithm is developed for fast approximation in case of large-scale problems.

Our model provides multiple advantages over other latent etiology estimation methods, such as Wu et al. (2015) and Wu et al. (2017). Most importantly, our model does not restrict the latent class to a pre-defined set of lung infection states. Rather, it automatically identifies the possible infection patterns out of all possibilities. Also, both the MCMC algorithm and the variational algorithm are quite scalable, and the scalability is especially important as the measurements are becoming more abundant.

A limitation of our method is that its estimation accuracy relies on the following assumptions. (1) The measurements are conditionally independent given the true latent status. (2) The experts' knowledge used for setting the TPR priors does not contradict with the truth. (3) The correlation structure of the latent nodes is sign-consistent, that is, their correlations are either all nonnegative or all nonpositive. Future work could be used to relax the above assumptions and make this method more robust. For example, we could borrow the nested structure proposed in Wu et al. (2017) to model the conditional dependence among measurements. We can modify the D matrix in the LSC model by adding a third state to allow for synergic effects of pathogens, but this could compromise the model identifiability. Regarding fast approximation, we have developed the variational inference algorithm based on the mean-field family (Blei et al., 2006), which could have a large bias when the posterior dependency between variables are strong, and could lead to significantly underestimated posterior variances. Thus, future work could involve explor-

ing different variational families that allow for interactions between factors, such as the partially factorized structure (Saul and Jordan, 1996) and the decimatable Boltzmann machine (Barber and Wiegerinck, 1999), so that the both the posterior mean and variance can be better approximated. Besides, since there is not much theory developed for the general asymptotic behavior of variational inference, an important extension of this work is to establish the theoretical guarantees of the variational approximation of the Bayesian latent sparse correlation model.

In Part II, we have developed efficient estimators for survival regression models by incorporating external information on the population level survival rates. The accelerated failure time (AFT) model and the Cox proportional hazards model are considered. For each model, over-identifying moment conditions are derived from the benchmark semi-parametric estimator (partial-likelihood estimator for Cox and log-rank estimator for AFT), and from the auxiliary survival information, by applying functional delta method. Parameter estimation and model diagnostics are carried out using the generalized method of moments (GMM) techniques. We have shown that the new GMM-based estimators are asymptotically and empirically more efficient than the benchmark estimators.

By proposing these estimators, we have introduced a novel approach for evidence synthesis in survival analysis with right-censoring data. Especially, the AFT-GMM estimators fill the gap of utilizing auxiliary information in the accelerated failure time model. Results show the efficiency improvement is significant when

the proportion of censoring is large or when auxiliary information is available for multiple time points. With the popular applications of the AFT model, our method is destined to have a significant impact on the practice of time-to-event analysis.

A1 Appendix to Chapter 5

Lemma 1

Let E be a $p-$dimensional positive definite matrix and D_0 be a $p-$dimensional symmetric matrix. Define O as a $q-$dimensional($q > p$) positive definite matrix, and D as a q by p matrix, such that

$$O = \begin{pmatrix} E & C^{*T} \\ C^* & A^* \end{pmatrix}, D = \begin{pmatrix} D_0 \\ D_1 \end{pmatrix}$$

where A^* is non-singular, then there exists a $(q-p)\times p$ matrix B and a $(q-p)\times(q-p)$ positive definite matrix A, such that $D^T O^{-1} D = D_0 E^{-1} D_0 + B^T A B$.

Proof: Given O, E, A^* are non-singular, we have

$$O^{-1} = \begin{pmatrix} A_1 & C^T \\ C & A_2 \end{pmatrix},$$

where

$$\begin{aligned} A_1 =&(E - C^{*T}A^{*-1}C^*)^{-1} \\ =&E^{-1} + E^{-1}C^{*T}A_2C^*E^{-1} \\ A_2 =&(A^* - C^*E^{-1}C^{*T})^{-1} \\ C^T =&- E^{-1}C^{*T}A_2. \end{aligned}$$

Therefore, define $J = C^*E^{-1}D_0$, $B = J - D_1$ and $A = A_2$, then

$$\begin{aligned} &D^TO^{-1}D \\ =&D_0E^{-1}D_0 + J^TA_2J - D_1^TA_2J - J^TA_2D_1 + D_1^TA_2D_1 \\ =&D_0E^{-1}D_0 + (J - D_1)^TA_2(J - D_1) \\ =&D_0E^{-1}D_0 + B^TAB \end{aligned}$$

Lemma 2

Let E be a $p-$dimensional positive definite matrix and D_0 be a $p-$dimensional symmetric matrix. Define O as a $q-$dimensional($q > p$) positive definite matrix, and

D as a q by p matrix, such that

$$O = \begin{pmatrix} E & C^{*T} \\ C^* & A^* \end{pmatrix}, D = \begin{pmatrix} D_0 & 0 \\ D_1 & D_2 \end{pmatrix},$$

where A^* is non-singular. Define F as the upper left $p \times p$ sub-matrix of $(D^T O^{-1} D)^{-1}$, then there exists a $(q-p) \times p$ matrix B and a $(q-p) \times (q-p)$ positive definite matrix A, such that $F = (D_0 E^{-1} D_0 + B^T A B)^{-1}$.

Proof: Given O, E, A^* are non-singular, we have

$$O^{-1} = \begin{pmatrix} A_1 & C^T \\ C & A_2 \end{pmatrix},$$

where

$$\begin{aligned} A_1 =& (E - C^{*T} A^{*-1} C^*)^{-1} \\ =& E^{-1} + E^{-1} C^{*T} A_2 C^* E^{-1} \\ A_2 =& (A^* - C^* E^{-1} C^{*T})^{-1} \\ C^T =& - E^{-1} C^{*T} A_2. \end{aligned}$$

Then

$$(D^TO^{-1}D)^{-1} = \begin{pmatrix} D_0A_1D_0 - D_1^TCD_0 - D_0C^TD_1 + D_1^TA_2D_1 & D_0C^TD_2 + D_1^TA_2D_2 \\ D_2^TCD_0 + D_2^TA_2D_1 & D_2^TA_2D_2 \end{pmatrix}^{-1}.$$

Therefore, define $J = C^*E^{-1}D_0$, $B = J - D_1$ and $A = A_2 - A_2D_2(D_2^TA_2D_2)^{-1}D_2^TA_2$, then we have

$$\begin{aligned} F =& \big[D_0A_1D_0 - D_1^TCD_0 - D_0C^TD_1 + D_1^TA_2D_1 - \\ & (D_0C^TD_2 + D_1^TA_2D_2)(D_2^TA_2D_2)^{-1}(D_2^TCD_0 + D_2^TA_2D_1)\big]^{-1} \\ =& \big[D_0E^{-1}D_0 + J^TA_2J - D_1^TA_2J - J^TA_2D_1 + D_1^TA_2D_1 - \\ & (J^TA_2D_2 + D_1^TA_2D_2)(D_2^TA_2D_2)^{-1}(D_2^TA_2J + D_2^TA_2D_1)\big]^{-1} \\ =& (D_0E^{-1}D_0 + B^TAB)^{-1} \end{aligned}$$

Proof of Theorem 2.1

Let $G_1 = \frac{1}{n}\sum_{i=1}^n g_i^{(1)}(\beta)$ and $G_2 = \big(\frac{1}{n}\sum_{i=1}^n g_{i1}^{(2)}(\beta), \ldots, \frac{1}{n}\sum_{i=1}^n g_{iK}^{(2)}(\beta)\big)$.

Let $D_0 = \mathbb{E}(\frac{\partial G_1}{\partial \beta}|\beta = \beta_0)$ and $\Sigma = \mathsf{var}(G_1|\beta = \beta_0)$.

Let $D_1 = \mathbb{E}(\frac{\partial G_2}{\partial \beta}|\beta = \beta_0)$

Let $C^* = \mathsf{Cov}(G_1, G_2|\beta = \beta_0)$ and $A^* = \mathsf{var}(G_2|\beta = \beta_0)$

Following the standard argument of GMM (**?**Newey and McFadden, 1994), it is

shown that the GMM estimate $\widehat{\beta}$ is consistent and the asymptotically normal with

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{D} N\Big(0, (D^T O^{-1} D)^{-1}\Big)$$

where

$$D = \begin{pmatrix} D_0 \\ D_1 \end{pmatrix}, O = \begin{pmatrix} \Sigma & C^{*T} \\ C^* & A^* \end{pmatrix}$$

By Lemma 1, we have $(D^T O^{-1} D)^{-1} = (D_0 \Sigma^{-1} D_0 + B^T A B)^{-1}$, where $B = C^* \Sigma^{-1} D_0 - D_1$ and $A = (A^* - C^* \Sigma^{-1} C^{*T})^{-1}$

Proof of Theorem 2.2 to 2.4

In the context of Theorem 2.2:

Define $G_1 = \frac{1}{n} \sum_{i=1}^n g_i^{(1)}(\beta)$, $G_2 = \big(\frac{1}{n} \sum_{i=1}^n g_{i1}^{(2)}(\beta, \rho), \ldots, \frac{1}{n} \sum_{i=1}^n g_{iK}^{(2)}(\beta, \rho)\big)$, and $\eta = \rho$.

In the context of Theorem 2.3:

Define $G_1 = \frac{1}{n} \sum_{i=1}^n h_i^{(1)}(\beta)$, $G_2 = \big(\frac{1}{n} \sum_{i=1}^n h_i^{(2)}(\beta, \alpha), \frac{1}{n} \sum_{i=1}^n h_{i1}^{(3)}(\beta, \alpha), \ldots, \frac{1}{n} \sum_{i=1}^n h_{iK}^{(2)}(\beta, \alpha)\big)$, and $\eta = \alpha$.

In the context of Theorem 2.4:

Define $G_1 = \frac{1}{n}\sum_{i=1}^{n} h_i^{(1)}(\beta)$, $G_2 = \big(\frac{1}{n}\sum_{i=1}^{n} h_i^{(2)}(\beta,\alpha), \frac{1}{n}\sum_{i=1}^{n} h_{i1}^{(3)}(\beta,\alpha,\rho), \ldots, \frac{1}{n}\sum_{i=1}^{n} h_{iK}^{(2)}(\beta,\alpha,\rho$ and $\eta = (\alpha, \rho)$.

Furthermore, define $\Sigma = \mathsf{var}(G_1|\beta = \beta_0)$, $D_0 = \mathbb{E}(\frac{\partial G_1}{\partial \beta}|\beta = \beta_0)$, $D_1 = \mathbb{E}(\frac{\partial G_2}{\partial \beta}|\beta = \beta_0, \eta = \eta_0)$, $D_2 = \mathbb{E}(\frac{\partial G_2}{\partial \eta}|\beta = \beta_0, \eta = \eta_0)$, $C^* = \mathsf{Cov}(G_1, G_2|\beta = \beta_0, \eta = \eta_0)$, and $A^* = \mathsf{var}(G_2|\beta = \beta_0, \eta = \eta_0)$. Following the standard argument of GMM (**?**Newey and McFadden, 1994), it is shown that the GMM estimate $(\widehat{\beta}, \widehat{\eta})$ is consistent and the asymptotically normal with

$$\sqrt{n}\Big(\begin{pmatrix} \widehat{\beta} \\ \widehat{\eta} \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \eta_0 \end{pmatrix} \Big) \xrightarrow{D} N\Big(0, (D^T O^{-1} D)^{-1}\Big),$$

where

$$D = \begin{pmatrix} D_0 & 0 \\ D_1 & D_2 \end{pmatrix}, O = \begin{pmatrix} \Sigma & C^{*T} \\ C^* & A^* \end{pmatrix}.$$

Let p be the length of β, and F be the upper left $p \times p$ block of $(D^T O^{-1} D)^{-1}$, then

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{D} N\Big(0, F\Big).$$

Finally, define $A_2 = (A^* - C^*\Sigma^{-1}C^{*T})^{-1}$, $A = A_2 - A_2 D_2 (D_2^T A_2 D_2)^{-1} D_2^T A_2$, and $B = C^*\Sigma^{-1}D_0 - D_1$. By Lemma 2, we have $F = (D_0\Sigma^{-1}D_0 + B^T A B)^{-1}$. Moreover,

in the context of Theorem 2.3 and 2.4, $D_0 = -\Sigma$, thus $F = (\Sigma + B^T AB)^{-1}$.

Bibliography

Adegbola, R. A. and Levine, O. S. (2011). Rationale and expectations of the pneumonia etiology research for child health (perch) study. *Expert review of respiratory medicine*, 5(6):731.

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society. Series A (General)*, 144(4):419–461.

Albert, P. S., McShane, L. M., Shih, J. H., Network, U. N. C. I. B. T. M., et al. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics*, pages 610–619.

Anatolyev, S. (2005). Gmm, gel, serial correlation, and asymptotic bias. *Econometrica*, 73(3):983–1002.

Ashley, E. A. (2015). The precision medicine initiative: a new national effort. *Jama*, 313(21):2119–2120.

BIBLIOGRAPHY

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. *Studies in item analysis and prediction*, 6:158–168.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Barber, D. and Wiegerinck, W. (1999). Tractable variational structures for approximating graphical models. In *Advances in Neural Information Processing Systems*, pages 183–189.

Berzofsky, M. and Biemer, P. P. (2012). Weak identifiability in latent class analysis. In *Proceedings of the ASA Section on Survey Methodology, Joint Statistical Meetings*, pages 4346–4354.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., Stuart, A., et al. (2013). Optimal tuning of the hybrid monte carlo algorithm. *Bernoulli*, 19(5A):1501–1534.

Bhattacharya, A. and Dunson, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association*, 107(497):362–377.

Bhattachayya, A. (1943). On a measure of divergence between two statistical population defined by their population distributions. *Bulletin Calcutta Mathematical Society*, 35(99-109):28.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Black, R. E., Cousens, S., Johnson, H. L., Lawn, J. E., Rudan, I., Bassani, D. G., Jha, P., Campbell, H., Walker, C. F., Cibulskis, R., et al. (2010). Global, regional, and national causes of child mortality in 2008: a systematic analysis. *The lancet*, 375(9730):1969–1987.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Blei, D. M., Jordan, M. I., et al. (2006). Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143.

Bou-Rabee, N. and Hairer, M. (2012). Nonasymptotic mixing of the mala algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110.

Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of

discrete choice. *Journal of the American Statistical Association*, 105(489):324–335.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press.

Calder, D. and Qazi, S. (2009). Evidence behind the who guidelines: hospital care for children: what is the aetiology of pneumonia in hiv-infected children in developing countries? *Journal of tropical pediatrics*, 55(4):219–224.

Cameron, J. L., Riall, T. S., Coleman, J., and Belcher, K. A. (2006). One thousand consecutive pancreaticoduodenectomies. *Annals of surgery*, 244(1):10.

Chatterjee, N., Chen, Y.-H., Maas, P., and Carroll, R. J. (2015). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, (just-accepted):1–32.

Chen, X. and Liao, Z. (2015). Sieve semiparametric two-step gmm under weak dependence. *Journal of Econometrics*, 189(1):163–186.

Clive, J., Woodbury, M. A., and Siegler, I. C. (1983). Fuzzy and crisp set-theoretic-based classification of health and disease. *Journal of Medical Systems*, 7(4):317–332.

Costantino, J. P., Gail, M. H., Pee, D., Anderson, S., Redmond, C. K., Benichou,

J., and Wieand, H. S. (1999). Validation studies for models projecting the risk of invasive and total breast cancer incidence. *Journal of the National Cancer Institute*, 91(18):1541–1548.

Cox, D. R. (1972). The analysis of multivariate binary data. *Applied statistics*, pages 113–120.

David, C. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34:187–220.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401):173–178.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278.

Deloria-Knoll, M., Feikin, D. R., Scott, J. A. G., OBrien, K. L., DeLuca, A. N., Driscoll, A. J., Levine, O. S., et al. (2012). Identification and selection of cases and controls in the pneumonia etiology research for child health project. *Clinical infectious diseases*, 54(suppl 2):S117–S123.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444.

Dunson, D. B. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Ekholm, A., McDonald, J. W., and Smith, P. W. (2000). Association models for a multivariate binary response. *Biometrics*, pages 712–718.

Ekholm, A., Smith, P. W., and McDonald, J. W. (1995). Marginal regression analysis of a multivariate binary response. *Biometrika*, 82(4):847–854.

Erosheva, E. A. (2006). Latent class representation of the grade of membership model. *Seatle: University of Washington*.

Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics*, 1(2):346–384.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80(1):141–151.

Franz, A., Adams, O., Willems, R., Bonzel, L., Neuhausen, N., Schweizer-Krantz, S., Ruggeberg, J. U., Willers, R., Henrich, B., Schroten, H., et al. (2010). Correlation of viral load of respiratory pathogens and co-infections with disease severity

in children hospitalized for lower respiratory tract infection. *Journal of clinical virology*, 48(4):239–245.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81(24):1879–1886.

Gelman, A. (2004). Parameterization and bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1996). Stochastic search variable selection. In *Markov chain Monte Carlo in practice*, pages 203–214. Springer.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. In *Markov Chain Monte Carlo in Practice*, pages 1–19. Springer.

Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamil-

tonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Guggenberger, P. (2008). Finite sample evidence suggesting a heavy tail problem of the generalized empirical likelihood estimator. *Econometric Reviews*, 27(4-6):526–541.

Gustafson, P. (2009). What are the limits of posterior distributions arising from non-identified models, and why should we care? *Journal of the American Statistical Association*, 104(488):1682–1695.

Gustafson, P., Le, N. D., and Vallée, M. (2002). A bayesian approach to case–control studies with errors in covariables. *Biostatistics*, 3(2):229–243.

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). Dram: efficient adaptive mcmc. *Statistics and Computing*, 16(4):339–354.

Haario, H., Saksman, E., Tamminen, J., et al. (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242.

Haberman, S. J. (1973). Log-linear models for frequency data: Sufficient statistics and likelihood equations. *The Annals of Statistics*, pages 617–632.

Haberman, S. J. (1995). Book review of statistical applications using fuzzy sets, by Kenneth G. Manton, Max A. Woodbury, and H. Dennis Tolley. *Journal of the American Statistical Association*, 90:1131–1133.

Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, pages 325–346.

Hammitt, L. L., Murdoch, D. R., Scott, J. A. G., Driscoll, A., Karron, R. A., Levine, O. S., OBrien, K. L., et al. (2012). Specimen collection for the diagnosis of pediatric pneumonia. *Clinical infectious diseases*, 54(suppl 2):S132–S139.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hirama, T., Yamaguchi, T., Miyazawa, H., Tanaka, T., Hashikita, G., Kishi, E., Tachi, Y., Takahashi, S., Kodama, K., Egashira, H., et al. (2011). Prediction of the

pathogens that are the cause of pneumonia by the battlefield hypothesis. *PloS one*, 6(9):e24474.

Horner, M., Ries, L., Krapcho, M., Neyman, N., Aminou, R., Howlader, N., Altekruse, S., Feuer, E., Huang, L., Mariotto, A., et al. (2015). Seer cancer statistics review, 1975-2015, national cancer institute. bethesda, md.

Huang, C.-Y., Qin, J., and Tsai, H.-T. (2015). Efficient estimation of the cox model with auxiliary subgroup survival information. *Journal of the American Statistical Association*, (just-accepted):00–00.

Huang, G.-H. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69(1):5–32.

Hussain, M., Tangen, C. M., Berry, D. L., Higano, C. S., Crawford, E. D., Liu, G., Wilding, G., Prescott, S., Kanaga Sundaram, S., Small, E. J., et al. (2013). Intermittent versus continuous androgen deprivation in prostate cancer. *New England Journal of Medicine*, 368(14):1314–1325.

Imbens, G. W. (2012). Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics*.

Imbens, G. W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61(4):655–680.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258.

Jin, Z., Lin, D., Wei, L., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353.

Jin, Z., Lin, D., and Ying, Z. (2006). On least-squares regression with censored data. *Biometrika*, 93(1):147–161.

Jones, G., Johnson, W. O., Hanson, T. E., and Christensen, R. (2010). Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics*, 66(3):855–863.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Kim, Y.-D. and Choi, S. (2007). Nonnegative tucker decomposition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Knott, M. and Bartholomew, D. J. (1999). *Latent variable models and factor analysis*. Number 7. Edward Arnold.

Korppi, M., Leinonen, M., Koskela, M., MÄkelÄ, P. H., and Launiala, K. (1989).

Bacterial coinfection in children hospitalized with respiratory syncytial virus infections. *The Pediatric infectious disease journal*, 8(10):687–691.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Levine, O. S., OBrien, K. L., Deloria-Knoll, M., Murdoch, D. R., Feikin, D. R., DeLuca, A. N., Driscoll, A. J., Baggett, H. C., Brooks, W. A., Howie, S. R., et al. (2012). The pneumonia etiology research for child health project: a 21st century childhood pneumonia etiology study. *Clinical infectious diseases*, 54(suppl 2):S93–S101.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, pages 13–22.

Lindley, D. V. (1972). *Bayesian statistics: A review*. SIAM.

Lipsitz, S. R., Fitzmaurice, G. M., Sleeper, L., and Zhao, L. (1995). Estimation methods for the joint distribution of repeated binary observations. *Biometrics*, pages 562–570.

Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J. E., Cousens, S., Mathers, C., and Black, R. E. (2016). Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the sustainable development goals. *The Lancet*, 388(10063):3027–3035.

Manton, K. G., Tolley, H. D., and Woodbury, M. A. (1994). *Statistical applications using fuzzy sets*. New York: John Wiley & Sons, cop.

Marshall, T. and Roberts, G. (2012). An adaptive approach to langevin mcmc. *Statistics and Computing*, 22(5):1041–1057.

McCutcheon, A. L. (1987). *Latent class analysis*. Number 64. Sage.

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika*, 21(4):331–347.

Meric-Bernstam, F. and Mills, G. B. (2012). Overcoming implementation challenges of personalized cancer therapy. *Nature reviews Clinical oncology*, 9(9):542–548.

Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.

Murdoch, D. R., OBrien, K. L., Driscoll, A. J., Karron, R. A., Bhat, N., et al. (2012). Laboratory methods for determining pneumonia etiology in children. *Clinical infectious diseases*, 54(suppl 2):S146–S152.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11).

Newey, W. K. and McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. volume 4, pages 2111–2245. Elsevier.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.

Newey, W. K. and West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, pages 777–787.

Owen, A. B. (2001). *Empirical likelihood*. Wiley Online Library.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, pages 1027–1057.

Park, Y. and Wei, L. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika*, pages 717–723.

Penny, W., Kiebel, S., and Friston, K. (2003). Variational bayesian inference for fmri time series. *NeuroImage*, 19(3):727–741.

Pericone, C. D., Overweg, K., Hermans, P. W., and Weiser, J. N. (2000). Inhibitory and bactericidal effects of hydrogen peroxide production by streptococcus pneumoniae on other inhabitants of the upper respiratory tract. *Infection and immunity*, 68(7):3990–3997.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Qin, J. (2000). Miscellanea. combining parametric and empirical likelihoods. *Biometrika*, 87(2):484–490.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, pages 300–325.

Qin, J., Zhang, H., Li, P., Albanes, D., and Yu, K. (2014). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika*, 102(1):169–180.

Regev-Yochay, G., Dagan, R., Raz, M., Carmeli, Y., Shainberg, B., Derazne, E., Rahav, G., and Rubinstein, E. (2004). Association between carriage of streptococcus pneumoniae and staphylococcus aureus in children. *Jama*, 292(6):716–720.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*. Springer.

Roberts, G. O. (1998). Optimal metropolis algorithms for product measures on the vertices of a hypercube. *Stochastics and Stochastic Reports*, 62(3-4):275–283.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268.

Roberts, G. O., Rosenthal, J. S., et al. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367.

Roberts, G. O., Rosenthal, J. S., et al. (2004). General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Saul, L. K. and Jordan, M. I. (1996). Exploiting tractable substructures in intractable networks. In *Advances in neural information processing systems*, pages 486–492.

Shang, W. and Wang, X. (2017). The generalized moment estimation of the additive–multiplicative hazard model with auxiliary survival information. *Computational Statistics & Data Analysis*, 112:154–169.

Shann, F. (1986). Etiology of severe pneumonia in children in developing countries. *The Pediatric Infectious Disease Journal*, 5(2):247–252.

Singh, V. and Aneja, S. (2011). Pneumonia–management in the developing world. *Paediatric respiratory reviews*, 12(1):52–59.

Singleton, R. J., Bulkow, L. R., Miernyk, K., DeByle, C., Pruitt, L., Hummel, K. B.,

Bruden, D., Englund, J. A., Anderson, L. J., Lucher, L., et al. (2010). Viral respiratory infections in hospitalized and community control children in alaska. *Journal of medical virology*, 82(7):1282–1290.

Stark, A. and Eibl, G. (2015). Pancreatic ductal adenocarcinoma. *Pancreapedia: The Exocrine Pancreas Knowledge Base*.

Sullivan, P. F., Kessler, R. C., and Kendler, K. S. (1998). Latent class analysis of lifetime depressive symptoms in the national comorbidity survey. *American Journal of Psychiatry*, 155(10):1398–1406.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate bayesian computation. *PLoS computational biology*, 9(1):e1002803.

Tikhomirova, A. and Kidd, S. P. (2013). Haemophilus influenzae and streptococcus pneumoniae: living together in a biofilm. *Pathogens and disease*, 69(2):114–126.

UNICEF, W., UNICEF, W., et al. (2006). Pneumonia: the forgotten killer of children. *UNICEF/WHO*, pages 1–40.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press.

Varga, J., Denton, C. P., Wigley, F. M., Allanore, Y., and Kuwana, M. (2016). *Scleroderma: From pathogenesis to comprehensive management*. Springer.

Vihola, M. (2012). Robust adaptive metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5):997–1008.

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, B., Titterington, D., et al. (2006). Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650.

Wang, Z., Mohamed, S., and Freitas, N. (2013). Adaptive hamiltonian and riemann manifold monte carlo. In *International Conference on Machine Learning*, pages 1462–1470.

Woodbury, M. A., Clive, J., and Garson Jr, A. (1978). Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and Biomedical Research*, 11(3):277–298.

World Health Organization et al. (1990). Programme for the control of acute respiratory infections. *Acute respiratory infections in children: case management in*

small hospitals in developing countries. A manual for doctors and other senior health workers. Geneva, Switzerland: World Health Organization.

World Health Organization, UNICEF, et al. (2009). Global action plan for prevention and control of pneumonia (gapp).

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193.

Wu, Z., Deloria-Knoll, M., Hammitt, L. L., and Zeger, S. L. (2015). Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Wu, Z., Deloria-Knoll, M., and Zeger, S. L. (2017). Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics*, 18(2):200–213.

Xing, E. P., Jordan, M. I., and Russell, S. (2002). A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 583–591. Morgan Kaufmann Publishers Inc.

You, C., Ormerod, J. T., and Müller, S. (2014). On variational bayes estimation and

variational information criteria for linear regression models. *Australian & New Zealand Journal of Statistics*, 56(1):73–87.

Young, M. A. (1983). Evaluating diagnostic criteria: a latent class paradigm. *Journal of Psychiatric Research*, 17(3):285–296.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, 77(3):642–648.

Zhou, M. (2006). The cox proportional hazards model with partially known baseline. *Random Walk, Sequential Analysis and Related Topics World Scientific*.

CURRICULUM VITAE

DETIAN DENG

ddeng3@jhu.edu

615 N. Wolfe St. E3037

Baltimore, MD 21205

Date of Birth: June 3rd, 1989

Place of Birth: Beijing, China

EDUCATION

2013 - 2018 **Johns Hopkins Bloomberg School of Public Health**, Baltimore, MD

Ph.D. in Biostatistics

Thesis title: *Scalable Semi-parametric Methods in Biostatistics*

Advisor: Prof. Scott L. Zeger

2017 **Johns Hopkins University**, Baltimore, MD

M.S.E in Computer Science

2011 - 2013 **Johns Hopkins Bloomberg School of Public Health**, Baltimore, MD

Sc.M. in Biostatistics

2007 - 2011 **Shanghai Jiao Tong University**, Beijing, China

---

CURRICULUM VITAE

B.S. in Biotechnology

PROFESSIONAL EXPERIENCE

05/2017 - 08/2017 **Data Scientist Intern**

Facebook, Menlo Park, CA

06/2016 - 08/2016 **Quantitative Analyst Intern**

Google, Mountain View, CA

HONORS AND AWARDS

JOHNS HOPKINS UNIVERSITY

2017 Finalist and Travel Award in Student Research Competition of American Public Health Association, Statistics Section

2015 Top Performer Award in DREAM Challenge: Prostate Cancer Prognosis

---

CURRICULUM VITAE

PUBLICATIONS

PUBLISHED/SUBMITTED

Deng, D. et al. Predicting survival time for metastatic castration resistant prostate cancer: An iterative imputation approach. F1000Research 2016, 5:2672

McCurdy, M., Bellows, A., **Deng, D.**, et al, Test-Retest Reliability of the Capute Scales in a High Risk Sample, Journal of Neonatal-Perinatal Medicine, vol. 8, no. 3, pp. 233-241, 2015

WORKING PAPERS

Deng, D., Zeger, SL., Bayesian Methods for Learning Multivariate Binary Hidden States with Application to Pneumonia Etiology Study

Deng, D., Huang, CY., Efficient Time-to-event Regression Estimation Incorporating Auxiliary Information with Generalized Method of Moments

---

PRESENTATIONS

2017 Bayesian Latent Variable Model with Sparse Correlation for Pneumonia Etiology Estimation . APHA annual meeting, Atlanta, GA

2016 Bayesian Marginal Regression Method with Latent Multivariate Binary Variables for Etiology Estimation. ENAR meeting, Austin, TX

2015 Bayesian Methods for Learning Multivariate Binary Hidden Variables. Amazon Graduate Research Symposium, Seattle, WA

TEACHING

2017 Practice of Statistical Consulting, 140.643.

2017 Statistical Machine Learning: Methods, Theory, and Applications, 140.644.

2016 Advanced Data Science **I-II**, 140.711-712.

2015 Statistical Methods in Public Health **I-III**, 140.621-623.

2014 Essentials of Probability and Statistical Inference **I-IV**, 140.646-649